MicroArray Explorer - A Java-based Tool For Data Mining Microarrays

Peter F. Lemkin[(1)], Gregory Thornwall[(2)] , Lothar Hennighausen[(3)]

[1]LECB, NCI/FCRDC, Frederick, MD 21702 (lemkin@ncifcrf.gov);
[2]SAIC/FCRDC; [3]LGP, NIDDK, Bethesda, MD

This paper was presented at the AMS-IMS-SIAM Summer Conference on Statistics in Functional Genomics (June 10-14, 2001) as "Data Mining Microarrays Using the MicroArray Explorer".

Revised: 09-28-2001

Abstract

The Microarray Explorer (MAExplorer) is a versatile Java-based data mining bioinformatic tool for analyzing quantitative cDNA expression profiles across multiple microarray platforms and DNA labeling systems. It may be run as either a stand-alone application or as a Web browser applet over the Internet. With this program it is possible to: 1) analyze the expression of individual genes; 2) analyze the expression of gene families and clusters; 3) compare expression patterns; 4) directly access other genomic databases for genes of interest. Data may be downloaded as required from a Web server or in the case of the stand-alone version, reside on the user's computer. Analyses are performed in real-time and may be viewed and directly manipulated in array pseudo-images, scatter plots, reports, histograms, expression profile plots, and cluster analyses plots. The raw data may be normalized under a variety of methods. A key feature is the gene data filter for constraining a working set of genes to those passing the intersection of a variety of user-specified logical and statistical tests. Resulting sets of genes may be saved as named sets and subsequent set operations may be performed resulting in derived gene sets. These in turn may be used in redefining the data filter. Similar sets of hybridized samples may be saved as named sets and used for reconfiguring experiment subsets. Reports may be generated with hypertext Web access to LocusLink, UniGene, GenBank, and other Internet genomic databases for sets of genes found to be of interest. Users may save their exploration states on the local computer, and later recall or share them with other scientists. The emphasis on direct manipulation of genes and sets of genes in graphics and tables provides a high level of interaction with the data, making it easier for investigators to test ideas when looking for patterns. A data format converter, Cvt2Mae, is available to convert data for use with MAExplorer. MAExplorer may be accessed at **http://www.lecb.ncifcrf.gov/MAExplorer**.

1. Introduction

One view of data mining is the discovery of putative relevant patterns. MAExplorer is a flexible Java-based microarray data-mining tool. It may be used stand-alone on an investigator's computer or in their Web browser as an applet when interacting with a compatible microarray database Web server. MAExplorer was designed to handle multiple samples as well as ordered expression profiles of genes across all samples. It handles either intensity or ratio (e.g. Cy3/Cy5) quantified microarray data. The array image spot detection and quantification phase is handled by other software prior to data analysis by MAExplorer. Users may drill down to relevant sets of genes using data-filters that

define a working gene set by statistics, clustering, and gene set operations. MAExplorer makes extensive use of direct manipulation of data in graphics and spreadsheets to take advantage of human visual pattern detection capabilities. Finally, when a set of genes of interest is found, additional information may be obtained from genomic Web servers by clicking on points in plots and entries in reports.

We originally developed MAExplorer (lemkin et al. 2000) as a Java applet for the Mammary Genome Anatomy Program (MGAP) (http://mammary.nih.gov/). MGAP includes information on specific models, histology associated with mammary tissue from these mouse models, and gene expression for some of the data (http://www.lecb.ncifcrf.gov/mae). Expression data has been obtained for some of the mouse models for about 1700 clones on 1) normal mammary development C57B6; 2) Knock-outs: Stat5a (-,-) and (+,-), CEBP null, TGF-Beta.; 3) Tumor models: WAP taq, p53 mutant, etc.; and 4) Transgenics: WAP Int3, BRCA1.

When used as a Web browser applet, each time MAExplorer is run it must download the applet and data from the Web server. This can be a slow and frustrating way to analyze the data. Subsequently we redeveloped MAExplorer as a more flexible stand-alone Java application that investigators may download and install on their computers avoiding these download delays each time they run it. This also makes it possible to use MAExplorer with an investigator's own array data as well as take advantage of other features only available with stand-alone applications.

The MAExplorer home page http://www.lecb.ncifcrf.gov/MAExplorer contains an extensive reference manual with many screen views illustrating its capabilities. It also contains tutorials, a demonstration database (from the MGAP database), a download area where users can freely download and install it for specific operating systems include Windows, MacOS, Solaris, Linux, Unix, and other Java compatible systems.

2. Methods

MAExplorer User Interface

The user interface is graphically oriented with a set of pull-down menus. We make heavy use of direct manipulation technology (Schneiderman, 1997) which lets users select spots in microarrays, points in various plots, and cells in reports to indicate data of interest. Selecting a particular gene declares it the "Current Gene". When clustering, and several clusters are involved, users indicate the current cluster by selecting a gene in one of the clusters. Figure 1 shows a typical analysis session consisting of the main MAExplorer window and additional pop up windows.

When used in stand-alone mode, MAExplorer is started by clicking on a user "Start.mae" file or by opening a file browser to a startup file for the selected database. This file was created either using the Cvt2Mae data format conversion program (to be discussed), a database server such as the NCI-CIT mAdb system (http://nciarray.nci.nih.gov) that generates the required set of files including the Start.mae file, or manually. The source data file schema and Cvt2Mae are described in the Reference Manual (Appendix C). Alternatively, one could manually edit a set of data files according to the schema to create a compatible set of database files.

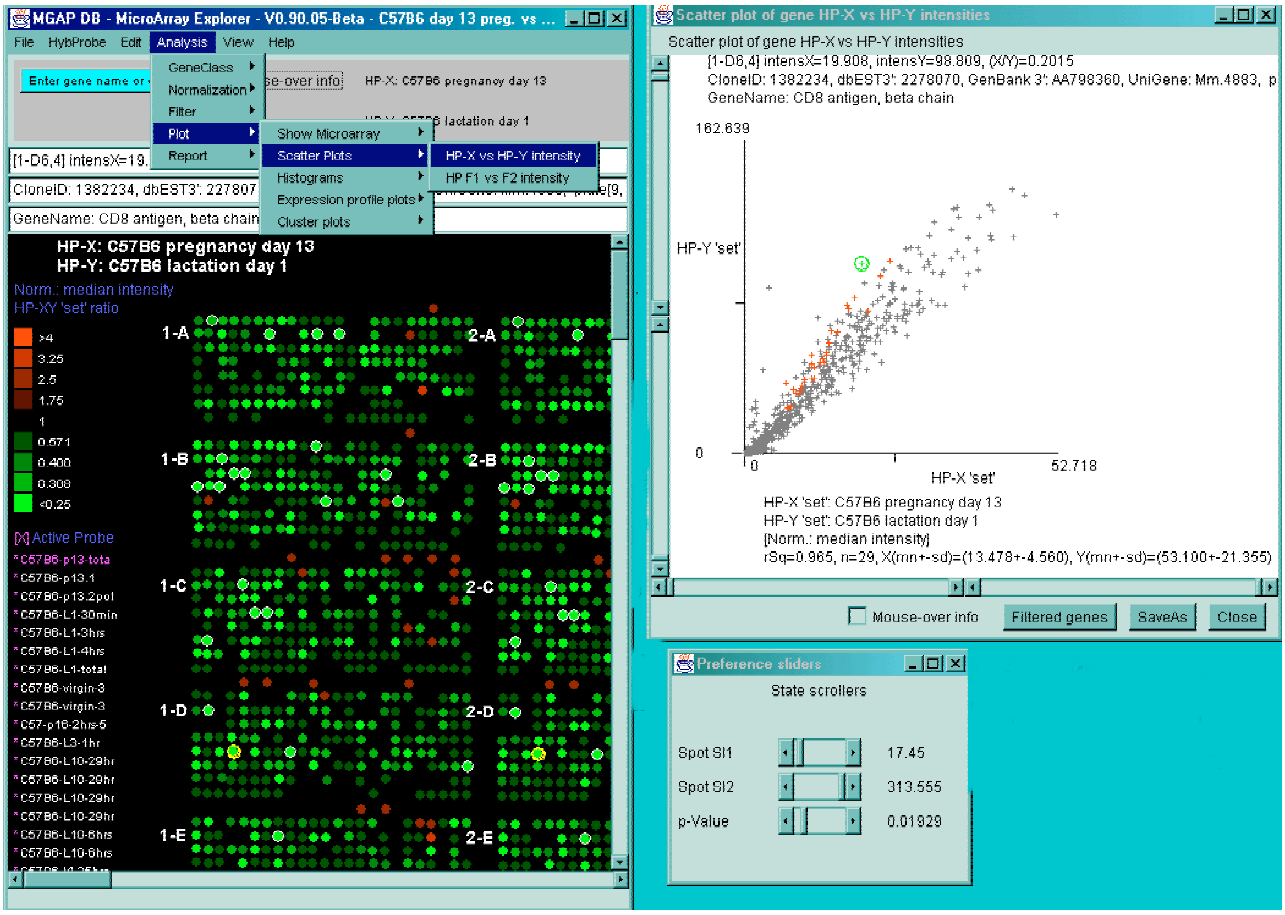

Figure 1. shows a typical MAExplorer data mining session. The data consists of 38 hybridized samples from the MGAP database of mouse mammary tissues. The X set is three C57B6 day 13 pregnancy and the Y samples are four C57B6 day 1 lactation samples. The pseudo array image in the main window shows the ratio of (X/Y), i.e. pregnancy/lactation according to the color chart on the left for a median normalization. The user interface shows the user selecting a scatter-plot (upper right) from the Plot pull-down submenu. Several data-filter threshold sliders are shown in the pop up State-Scrollers window (lower right). These thresholds are associated with the current data filters (Gene Class set to all named genes, spot intensity range [SI1:SI2] set to remove very low intensity noise data, and a low t-test [p-value] comparing the X and Y sets) resulting in 29 genes passing the filter. Finally, a scatter plot shows all of the genes in the database with those passing the data filter indicated in red and those failing the filter in gray. The scrollbars on the scatter plot let users zoom in on particular regions of interest. The current gene selected is show as a green circle in the scatter plot with annotation data at the top of the scatter plot and the top of the main window (left). The pseudo array image shows the current gene as a yellow circle and genes passing the data filter with white circles (difficult to see in the publication figure).

Source Data File Schema

The data file schema consists of a set of tab-delimited data files: including a Gene In Plate Order (GIPO or array print file); a list of samples in the user's database; a configuration file describing the array geometry, type of data, etc.; and finally a set of spot quantification data files – one for each hybridized sample. This architecture allows one to incrementally add new samples to an

existing database. Files are read from either a local disk (in the case of the stand-alone application method) or from an array database of a coordinated Web server (the browser applet method). It may be used with any experiment data that uses this schema. In addition to the original Research Genetics arrays for MGAP, it has been used with membranes from NIA, Incyte Cy3/Cy5 data, Affymetrix data, and others. In collaboration with John Powell and Ester Asaki, we are able to download data sets using this schema from the NCI/CIT mAdb relational database Web server servicing the NIH community. The data conversion tool, Cvt2Mae, may be used to convert generic array data files (Incyte, Affymetrix, and others) to this schema as well as allowing the user to describe their own array data files.

Lists of Hybridized Samples

A database contains multiple hybridized array samples that may consist of many conditions - some with replicate samples. We organize these samples as lists of conditions. The three cases are: 1) pairs of individual samples of 2 conditions (called X and Y); 2) sets of replicates samples of 2 conditions (called the X-set and Y-set); 3) an ordered expression profile list (E-list) of any subset of database samples. There are menu operations to let the user graphically assign samples to the X-set, Y-set and E-lists. Users can further manage these sample lists by manipulating named lists of sample conditions and create new lists using Boolean operations (And, Or, Difference). As these lists are part of the exploratory data analysis session-state, they are saved when the session is saved as a new startup file. When MAExplorer is restarted at some future time using this saved startup file, the session is restored to the previous state.

Data Filters

A data filter is the intersection of one or more tests. Each test returns a list of genes that meet some criteria defined by that test. These include roughly four categories: 1) gene subset membership; 2) spot intensity and ratio ranges, 3) statistical tests CV, t-Test, etc.; and 4) clustering tests.

We define a gene set "data-filter" that is applied in stages as a pipeline of tests. This successively performs the gene set intersection of the previous tests in the pipeline. The initial set is the set of all genes. For example, a gene membership test could be for Gene Class membership (e.g. All named genes, or All ESTs, etc), or a "User Gene Set" defined by the user using various set operations (to be discussed). A working set of genes resulting from applying the data filter may be saved by the user as a named gene set and subsequently manipulated using set operations. Gene sets in turn may be used as part of subsequent data filter or for normalization operations. Since these sets are part of the data mining session they are saved when the session is saved. The working gene set is used as the gene set pre-filter for subsequent clustering, plots, and reports.

Many of the data filters have associated pop up threshold slider bars that may be adjusted directly to change the filter parameters. Settings are determined by the investigator using visual feedback from the graphical results to decide on appropriate values. Changing a slider automatically re-computes the data filter and updates all active plots and clustering operations.

Plots

Graphic plots allow visualization and direct manipulation of numeric and logical gene data. MAExplorer provides a variety of visualization methods. 1) The color-coded pseudo array image displays either intensity or ratio (X/Y) data for each spot in the array in a variety of intensity and ratio color-coded models (see Figure 1). 2) Zoomable scatter plots may be used for viewing individual X vs Y samples, Cy3 vs Cy5 channels of the same or different samples, duplicate spots or X-set vs Y-set replicate sample data. The zoom feature is useful when data points are in close proximity and can be separating by zooming in. When used with K-means clustering, clicking on a point in the display indicates (in green numbers) the genes belonging to that cluster (see Figure 2). Clicking on a point in the plot defines it as the current gene. If K-means clustering is active, it is used to define the current cluster (as those genes that are in the same cluster that the current gene is in). Information about the current gene is shown in the scatter plot and main pseudo array window (see Figure 1). If the genomic Web database access is activated (not shown), then selecting the current gene will pop up a Web browser with data from that genomic database for that gene. 3) Both ratio and intensity histograms may be made of various collections of data and may be used for data filtering by clicking on particular bins and associated range operators. For example, one could select a ratio and include or exclude genes in the symmetric range (eg. >3X and <0.333X). 4) Expression profile plots may be created for individual genes. These per-gene plots may be assembled into scrollable lists of expression profile plots and are most useful when there is an ordered list of genes (e.g. the set of similar genes). Expression profiles may also be displayed as overlay plots. 5) Silhouette plots are used with similarity clustering and with K-means clustering, and are useful for determining the variation within a cluster. 6) Hierarchical clustering may be used to create clustergrams (Weinstein etal. 1997) and dendrograms of the working set of genes. Examples of all of these plots and others may be found in the Reference Manual.

Data reports

Data may be reported in tables in two formats. The first is a Web-accessible dynamic spreadsheet that allows users to click on cells in the spreadsheet and have it bring up a Web browser window for the gene in the specified database column. These databases could be LocusLink, GenBank, UniGene, the NCI/CIT mAdb clone reports, etc. The second format is a tab-delimited text area that is easily exported to Excel using either cut and paste or saving it to a tab-delimited text file. There are a variety of sample reports and gene set reports. Examples include an upper-diagonal sample vs sample correlation table for the current data filtered genes, a set of the N genes with highest ratios, etc. These are detailed in the Reference Manual.

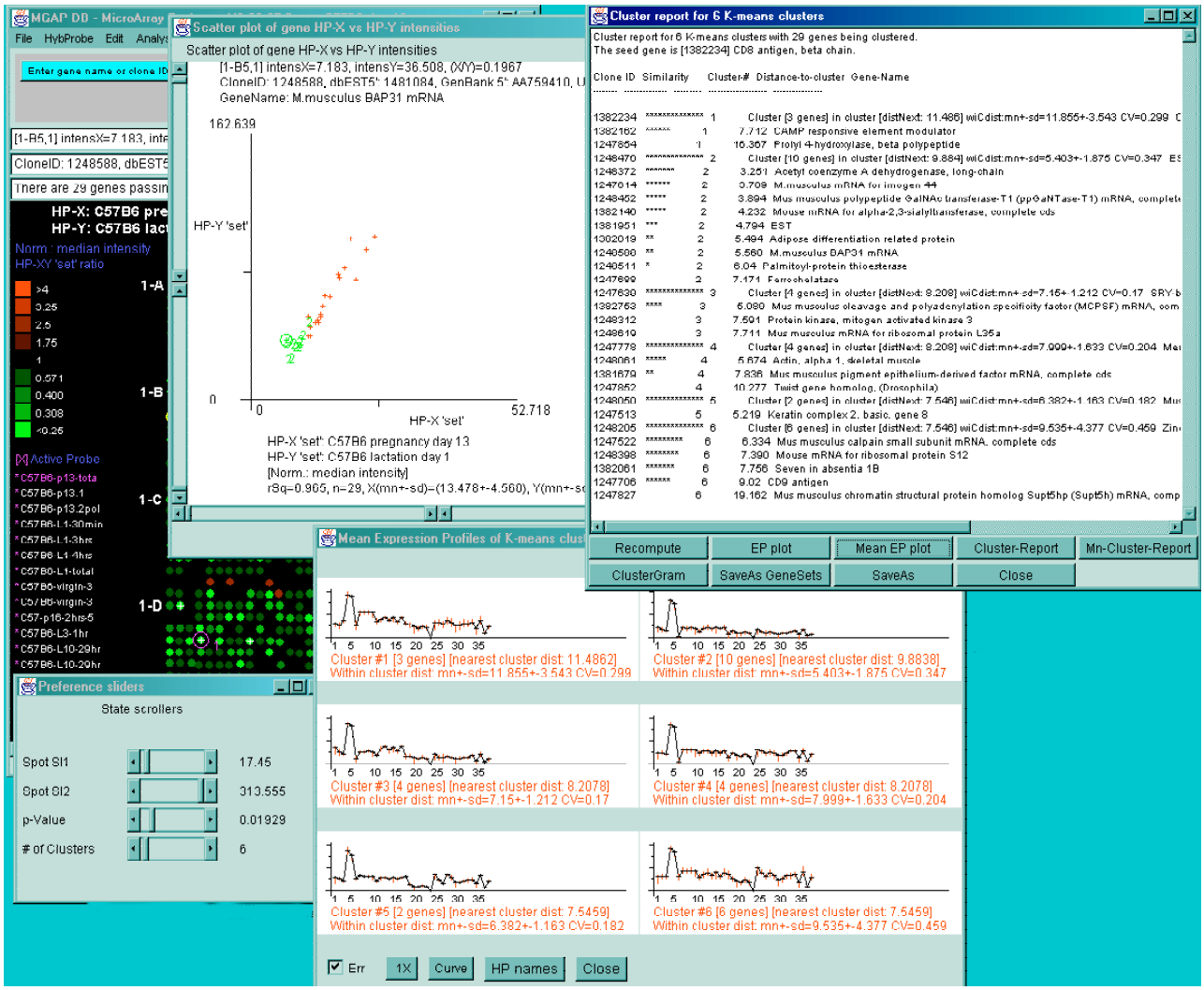

Figure 2. shows the results of applying K-means gene clustering to the pre-filtered set of 29 genes shown in Figure 1 for 6 clusters. The number of clusters is set by the "# of clusters" scrollbar (lower left). The scatter plot (upper left) has been set to show just the filtered genes and indicates the selected current cluster #2 as green number 2's. A cluster report (upper right) shows the similarity-ordered list of gene clusters with silhouette plots and gene names. Various options are available after the clustering and are shown in the command buttons at the bottom of the report. One of these, the mean expression profile plot (lower right), is shown at the bottom of the screen where the mean expression plots for each of the clusters are plotted. The expression profile was computed across all of the 38 samples using a Euclidian distance measure (the correlation coefficient could be used as an alternate distance measure).

Gene Set Operations

Because of the complexity of the data and the combinatoric nature of comparing many subsets of samples, it is useful to catalog gene subsets as named gene sets to help manage this complexity. There are a number of special gene sets that are part of the normal operation of the system as well as user defined saved gene sets or sets derived by Boolean operations (And, Or, Difference) between existing sets (see Figure 3). The special sets include the: 1) "Data Filtered genes" that

holds genes passing the data filter; 2) "Edited Gene List" holds results of gene clustering or editing operations; 3) "Normalization genes" the user defined gene set that may be used as one of the normalization method and 4) "User Gene Set" that may be used as one of the data filters. Specific genes or subsets of genes may be specified by name or using a wildcard name subset (e.g. "ONCO*" to find all oncogenes). The set of genes resulting from using the wildcard search is saved in the Edited Gene List and may be used for other operations. Performing this multiple times for different keywords and performing gene set union operations on the edited results lets the user define a set of ontology gene sets. All gene sets are saved when the session is saved, and restored when MAExplorer is restarted. So this may be useful for organizing the data.

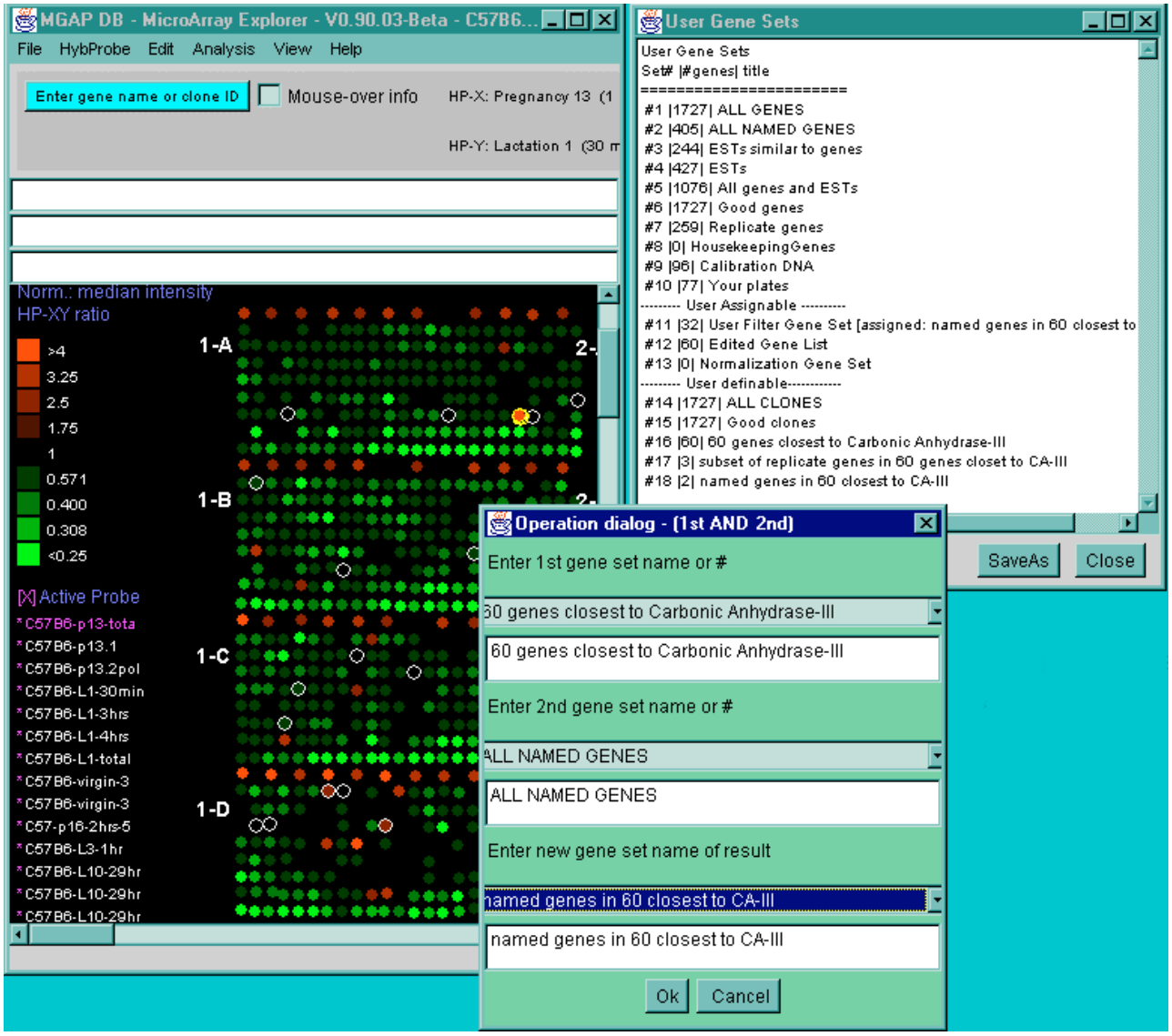

Figure 3. shows another data mining experiment on the same database illustrating some of the gene set operations. This time, a similarity cluster (not shown) was performed on the 38 samples to see which genes were closest to Carbonic Anhydrase-III (yellow circle in main window) from the set of all named genes and ESTs. The 60 most similar genes were selected and saved in a named gene set. The Boolean AND gene set operation (lower right) was performed on (a) the 60 similar genes, and (b) All-named-genes. The resulting subset was called "named genes in 60 closest to CA-III". The list of named gene sets is shown in the upper right.

Plug-in Extensions to MAExplorer

In order to open up MAExplorer to new investigator initiated analytic methods, we are in the process of adding a Java Plug-ins facility to MAExplorer. This will greatly extend the capabilities of the core functionality. We have added a Web page to the MAExplorer Web site where there will be information about the plug-ins including an open Java Application Programming Interface

(API), examples of MAExplorer Java plug-ins source code, a catalog of donated plug-ins. It will also include links to other Web sites that wish to do their own distribution of their plug-ins. Typical plug-ins might include new normalization methods, data filters, PCA or multi-dimensional scaling, new clustering methods, access to genomic Web-servers that may do additional functional analysis of cluster results, etc. Plug-ins are divided into three types: 1) using 100% Java so the plug-in is totally portable; 2) using a Java "stub" to access local programs written in any language (e.g. The "R" statistical package); and 3) Java "stubs" to access Web-CGI or client-server to connect to specialized genomic databases where additional functional analysis of gene subsets could be performed. Some plug-ins could allow MAExplorer to function as a front-end tool for back-end databases that offer other types of data and analysis methods.

Data format conversion for MAExplorer using Cvt2Mae

As was mentioned above, MAExplorer requires input files be in a particular data file schema. The Cvt2Mae tool reads a variety of array data - for both one-of-a kind academic and commercial arrays (eg. Affymetrix, Incyte, etc). It lets the user create an "array layout" description that may be used in subsequent conversions. We will be adding the standard XML data access using the "MGED" microarray data portability standard when it becomes available.

Summary

MAExplorer is a flexible data-mining tool for microarray data mining. It uses direct-manipulation, data filtering, built-in graphics, statistics, clustering, gene and sample set operations. Many of the facilities are used to manage multiple samples, and gene sets allowing the user to easily construct different views of the same data to possibly discover patterns appearing in some of the views. We are extending the design of the software to allow investigators to add their own analytic methods using Java plug-ins. We will make those we have access to publicly available and would encourage others to do so as well. MAExplorer and Cvt2Mae may be downloaded from http://www.lecb.ncifcrf.gov/MAExplorer where online documentation (manual, tutorials, etc.) is available.

Acknowledgements

We would like to thank Ester Asaki, Kevin Becker, Damien Chaussabel, Chris Cheadle, Terry Clark, Yongzhi Karen Cui, Jai Evans, Josef Jurrek, Mitko Dimitrov, John Powell, Jean-Pierre Renou, Chris Santos, Moshe Shani, Garry Smithers, Bob Stephens, Mark Vawter, Kathleen Walton, and many others for useful suggestions and work on the system.

References

1. Lemkin, P.F., Thornwall, G.C., Walton, K.D., Hennighausen, L. (2000 The Microarray Explorer tool for data mining of cDNA microarrays – application for the mammary gland. *Nucleic Acids* Res. **28**:4452-4459.

2. Schneiderman, B. (1997) *Designing the Human Interface*, 3rd edition. Addison-Wesley Pub. Co., NY, pp 1-638.

3. Weinstein J.N., Myers T.G., O'Conner, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johynson, G.S., Wittes, R.E., Paul, K.D. (1997) An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **275**: 343-349.
