ProtPlot – A Tissue Molecular Anatomy Program Java-based Data Mining Tool: Screen Shots

****** DRAFT - undergoing revision ******

Peter F. Lemkin, Ph.D.[1], Djamel Medjahed Ph.D.[2]
[1]NCI-Frederick; [2]SAIC-Frederick, MD 21702

Home page: http://tmap.sourceforge.net/

Revised: 08-26-2004 Version 0.40.1-*Beta*

Abstract

- ProtPlot is an open-source Java-based data mining bioinformatic tool for analyzing CGAP- database derived estimated mRNA tissue EST expression in terms of a set of virtual 2D-gels
- The estimated mRNA expression is mapped to estimated “proteins”
- It is well known, mRNA expression generally does *not* correlate well with protein expression as seen in 2D-PAGE gels (Ideker et.al., *Science* 292:929-934, 2001
- ProtPlot lets you look at the data in new ways and may help in thinking about new hypotheses for protein post-modifications or mRNA post-transcription processing.

Possible Questions

- **ProtPlot may help look at aggregates of CGAP data in new ways:**

  - Which "estimated proteins" are in a particular (pI,Mw) range?

  - Which sets of "proteins" are up or down regulated in cancer(s) and normal(s) or precancer(s)?

  - Which sets of "proteins" are entirely missing in one condition vs. the other?

  - Which sets of "proteins" cluster together across different types of cancers or normals?

ProtPlot

- It was developed initially as Virtual-2D [Proteomics J, in press], and upcoming paper on TMAP [Proteomics, in press]

- ProtPlot was derived from an open-source microarray data mining tool MAExplorer (http://maexplorer.sourceforge.net/) by P. Lemkin

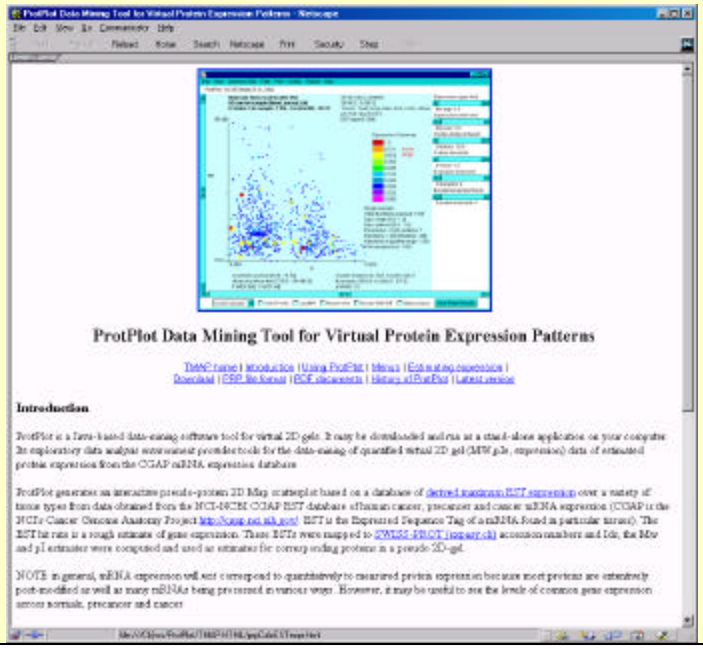
- ProtPlot is a Java application and runs on your computer. You download and install the application and the data.

Pseudo 2D-Gel Map Expression Data

- Sample mRNA estimated expression data was obtained for a variety of human tissue and histology types (normal, pre-cancer, cancer) using the relative hit rates on cDNA clone libraries. Data from multiple libraries/tissue were merged
- Pseudo-protein data was computed by mapping the UniGene Ids in the CGAP libraries to SwissProt AC. The (pI, Mw) was computed using the SwissProt (pI,Mw) server tool
- These data are assembled into ProtPlot data files called .prp files described on the Web site.
- ProtPlot then generates an interactive pseudo 2D-gel Map (pIe,Mw) scatterplot that may be used for data mining

VIRTUAL2D home: http://proteom.ncifcrf.gov/

TMAP HOME: http://tmap.sourceforge.net/



ProtPlot Data Mining Tool for Virtual Protein Expression Patterns

TMAP home | Introduction | Using ProtPlot | Menus | Estimating expression | Download | PRP file format | PDF documents | History of ProtPlot | Latest version

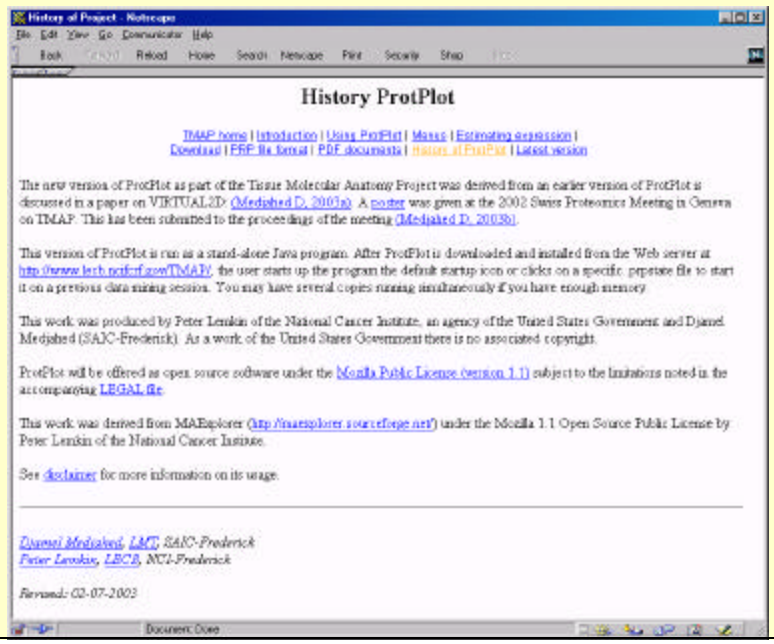
Introduction

ProtPlot is a Java-based data-mining software tool for virtual 2D gels. It may be downloaded and run as a stand-alone application on your computer. It's exploratory data analysis environment provides tools for the data-mining of quantified virtual 2D gel (MW,pI,e, expression) data of estimated protein expression from the CGAP mRNA expression database.

ProtPlot generates an interactive pseudo-protein 2D Map, scatterplot based on a database of derived transcript EST expression over a variety of tissue types from data obtained from the NCI-NICHE CGAP EST database of human cancer, precancer and cancer mRNA expression (CGAP is the NCI's Cancer Genome Anatomy Project http://cgap.nci.nih.gov/. EST is the Expressed Sequence Tag of mRNA found in particular tissues). The EST hit rate is a rough estimate of gene expression. These ESTs were mapped to SWISS-PROT (primary ch) accession numbers and Mr, the MW and pI estimates were computed and used as estimates for correctly ordering proteins in a pseudo 2D-gel.

NOTE in general, mRNA expression will not correspond to quantitatively to measured protein expression because most proteins are extensively post-modified as well as many mRNAs being processed in various ways. However, it may be useful to see the levels of common gene expression across normals, precancer and cancer.

History of ProtPlot



History ProtPlot

TMAP home | Introduction | Using ProtPlot | Menus | Estimating expression | Download | PRP file format | PDF documents | History of ProtPlot | Latest version

The new version of ProtPlot as part of the Tissue Molecular Anatomy Project was derived from an earlier version of ProtPlot as discussed in a paper on VIRTUAL2D: (Mediahed D, 2003a). A poster was given at the 2002 Swiss Proteomics Meeting in Geneva on TMAP. This has been submitted to the proceedings of the meeting (Mediahed D, 2003b).

This version of ProtPlot is run as a stand-alone Java program. After ProtPlot is downloaded and installed from the Web server at http://www.lecb.ncifcrf.gov/TMAP, the user starts up the program the default startup icon or clicks on a specific .prpstate file to start it on a previous data mining session. You may have several copies running simultaneously if you have enough memory.

This work was produced by Peter Lemkin of the National Cancer Institute, an agency of the United States Government and Djamel Mediahed (SAIC-Frederick). As a work of the United States Government there is no associated copyright.

ProtPlot will be offered as open source software under the Mozilla Public License (version 1.1) subject to the limitations noted in the accompanying LEGAL file.

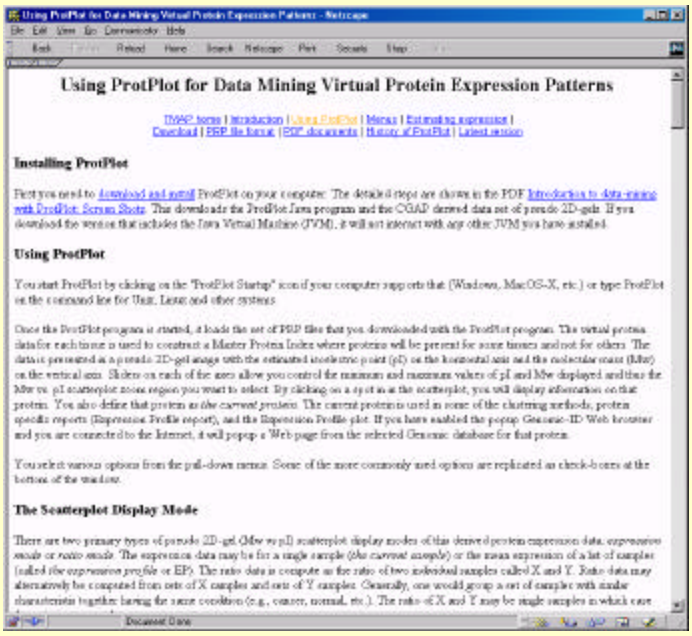
This work was derived from MAExplorer (http://maexplorer.sourceforge.net/) under the Mozilla 1.1 Open Source Public License by Peter Lemkin of the National Cancer Institute.

See disclaimer for more information on its usage.

Djamel Mediahed, L&T, SAIC-Frederick
Peter Lemkin, LECB, NCI-Frederick

Revised: 02-07-2003

Using ProtPlot



Using ProtPlot for Data Mining Virtual Protein Expression Patterns

TMAP home | Introduction | Using ProtPlot | Menus | Estimating expression | Download | PRP file format | PDF documents | History of ProtPlot | Latest version

Installing ProtPlot

First you need to download and install ProtPlot on your computer. The detailed steps are shown in the PDF Introduction to data mining with ProtPlot: Screen Shots. The downloads are the ProtPlot Java program and the CGAP derived data set of pseudo 2D-gels. If you download the version that includes the Java Virtual Machine (JVM), it will not interact with any other JVM you have installed.

Using ProtPlot

You start ProtPlot by clicking on the "ProtPlot Startup" icon if your computer supports that (Windows, MacOS-X, etc.) or type ProtPlot on the command line for Unix, Linux and other systems.

Once the ProtPlot program is started, it loads the set of PRP files that you downloaded with the ProtPlot program. The virtual protein data for each tissue is used to construct a Master Protein Index where proteins will be present for some tissues and not for others. The data is presented as a pseudo 2D-gel image with the estimated isoelectric point (pI) on the horizontal axis and the molecular mass (kDa) on the vertical axis. Sliders on each of the axes allow you control the minimum and maximum values of pI and Mw displayed and thus the Mw vs. pI scatterplot zoom region you want to select. By clicking on a spot in the scatterplot, you will display information on that protein. You also define that protein as *the current protein*. The current protein is used in some of the clustering methods, protein specific reports (Expression Profile report), and the Expression Profile plot. If you have enabled the popup Genomic-ID Web browser and you are connected to the Internet, it will popup a Web page from the selected Genomic database for that protein.

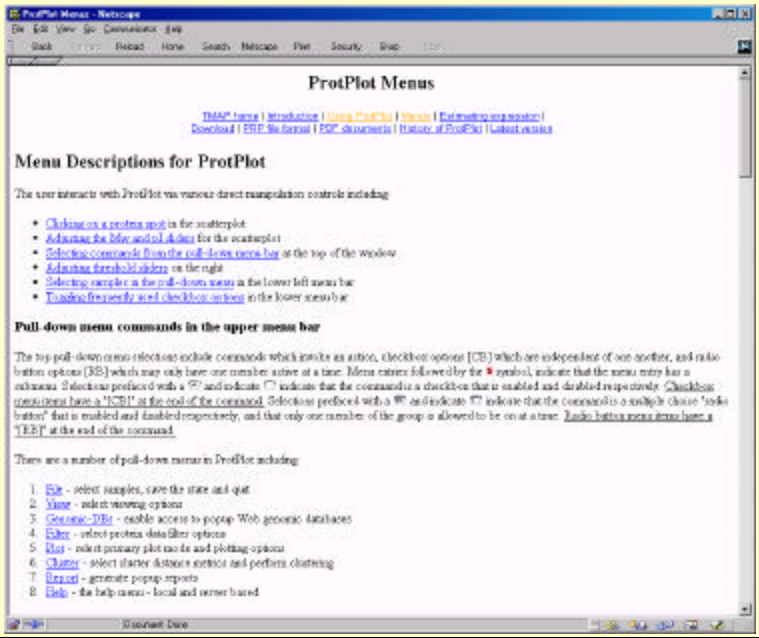
You select various options from the pull-down menus. Some of the more commonly used options are replicated as check-boxes at the bottom of the window.

The Scatterplot Display Mode

There are two primary types of pseudo 2D-gel (Mw vs pI) scatterplot display modes of the derived protein expression data, *expression mode* or *ratio mode*. The expression data may be for a single sample (*the current sample*) or the mean expression of a list of samples (called *the expression profile* or EP). The ratio data is computed as the ratio of two individual samples called X and Y. Ratio data may alternatively be computed from sets of X samples and sets of Y samples. Generally, one would group a set of samples with similar characteristics together, having the same condition (e.g., cancer, normal, etc.). The ratio of X and Y may be single samples in which case

Document Done

ProtPlot Menus and User Controls



ProtPlot Menus

TMAP home | Introduction | Using ProtPlot | Menus | Estimating expression | Download | PRP file format | PDF documents | History of ProtPlot | Latest version

Menu Descriptions for ProtPlot

The user interacts with ProtPlot via various direct manipulation controls including:

- Clicking on a protein spot in the scatterplot
- Adjusting the Mw and pI sliders for the scatterplot
- Selecting commands from the pull-down menu bar at the top of the window
- Adjusting threshold sliders on the right
- Selecting samples in the pull-down menu in the lower left menu bar
- Toggling frequently used checkbox options in the lower menu bar

Pull-down menu commands in the upper menu bar

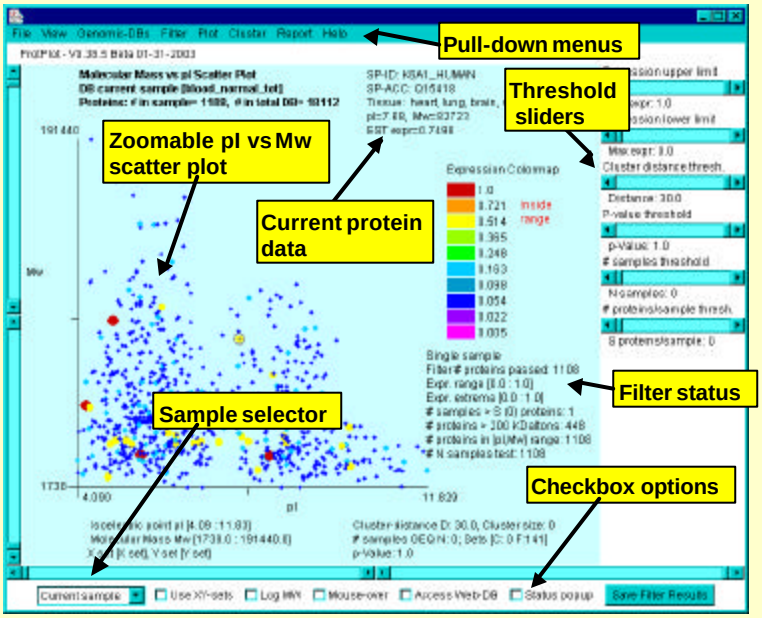
The top pull-down menu selections include commands which invoke an action, checkbox options [CB] which are independent of one another, and radio button options [RB] which may only have one member active at a time. Menu entries followed by the ▼ symbol, indicate that the menu entry has a submenu. Selections prefaced with a ☑ and indicate ☐ indicate that the command is a checkbox that is enabled and disabled respectively. Checkbox commands have a "[CB]" at the end of the command. Selections prefaced with a ◉ as indicate ○ indicate that the command is a multiple choice "radio button" that is enabled and disabled respectively, and that only one member of the group is allowed to be on at a time. Radio button menu items have a "[RB]" at the end of the command.

There are a number of pull-down menus in ProtPlot including:

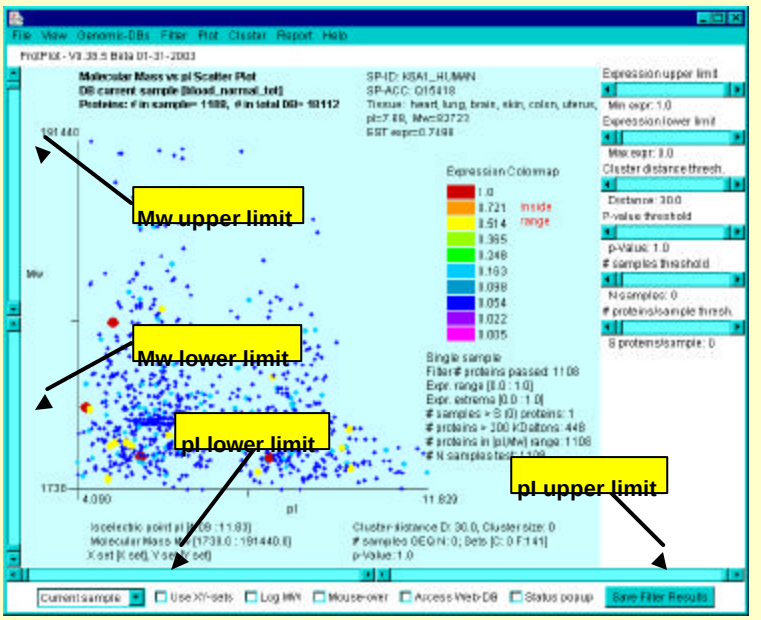
1. File - select samples, save the state and quit
2. View - select viewing options
3. Genomic-DBs - enable access to popup Web genomic databases
4. Filter - select protein data filter options
5. Plot - select primary plot modes and plotting options
6. Cluster - select cluster distance metrics and perform clustering
7. Report - generate popup reports
8. Help - the help menu - local and server based

Document Done

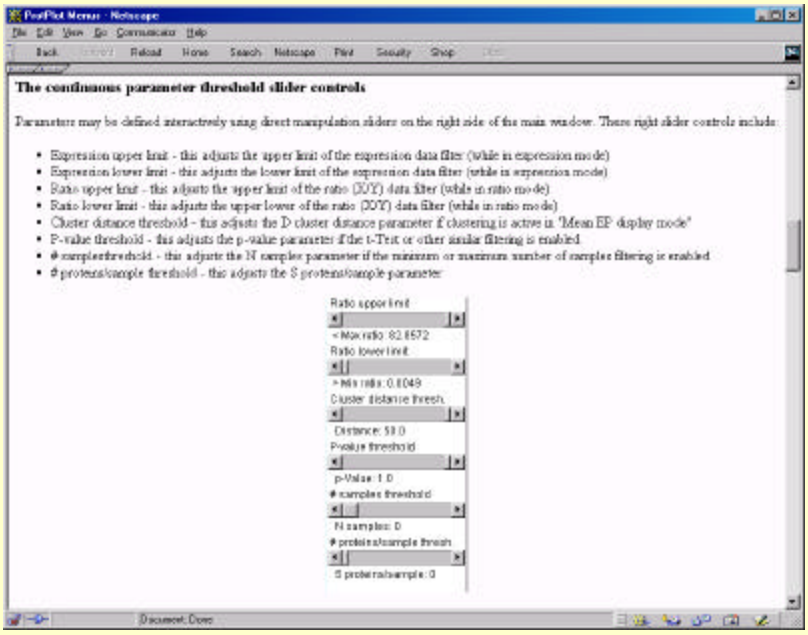
Initial Screen displaying (pI vs Mw) scatterplot



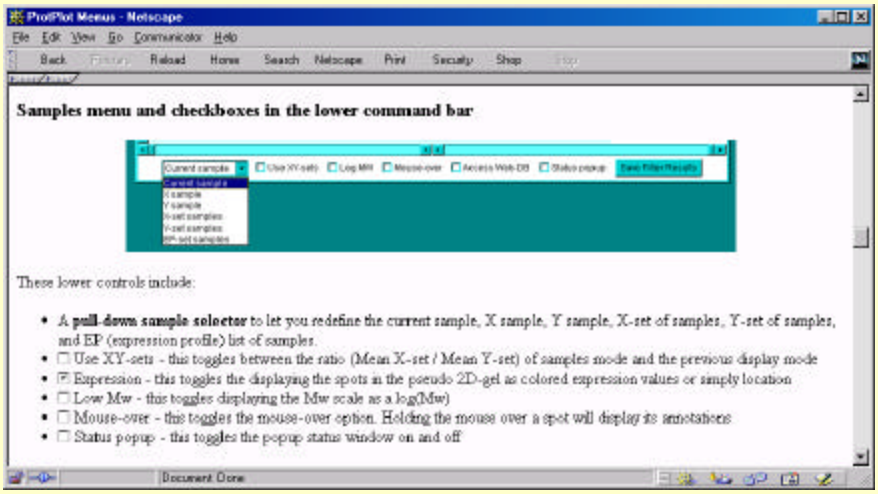
Scatterplot pI vs Mw Limit Sliders



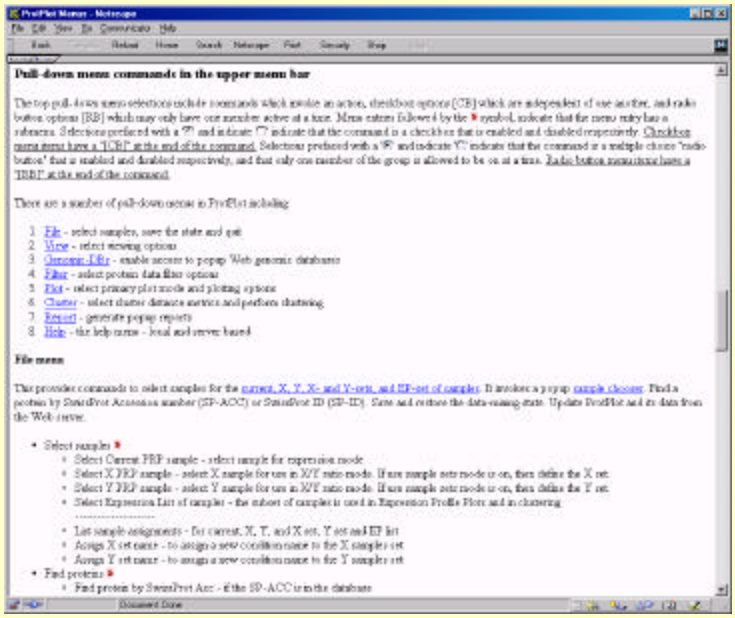
ProtPlot: Parameter Threshold Sliders



ProtPlot: Lower Selectors and Checkboxes



ProtPlot Pull-down Menus



Pull-down menu commands in the upper menu bar

The top pull-down menu selections include commands which invoke an action, checkbox options [CB] which are independent of one another, and radio button options [RB] which may only have one member active at a time. Menu entries followed by the ▶ symbol, indicate that the menu entry has a submenu. Selections prefaced with a ☐ and indicate ☐ indicate that the command is a checkbox that is enabled and disabled respectively. Checkbox menu items have a 'CHB' at the end of the command. Selections prefaced with a 🔘 and indicate '◯' indicate that the command is a multiple choice 'radio-button' that is enabled and disabled respectively, and that only one member of the group is allowed to be on at a time. Radio button menu items have a 'RBB' at the end of the command.

There are a number of pull-down menus in ProtPlot including:

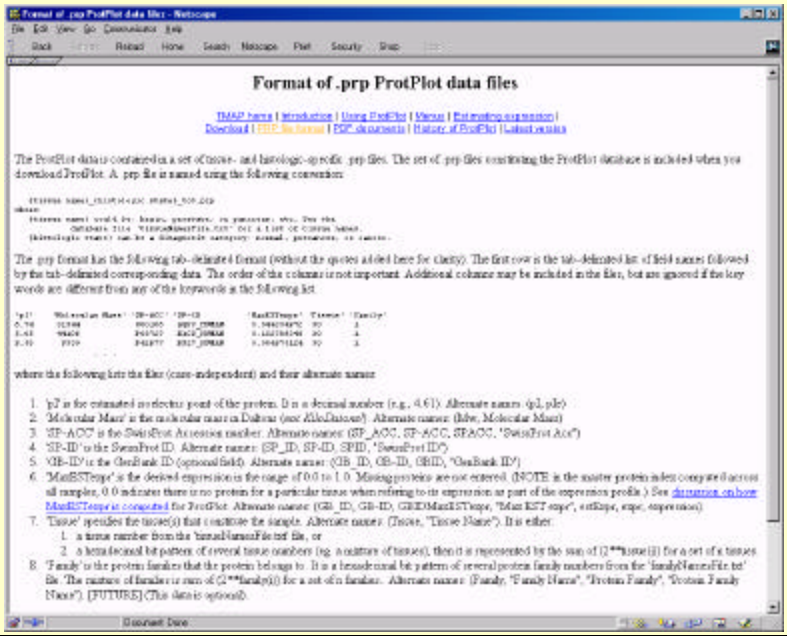
1. File - select samples, save the state and quit
2. View - select viewing options
3. Genome DBs - enable access to popup Web genomic databases
4. Filter - select protein data filter options
5. Plot - select primary plot mode and plotting options
6. Cluster - select cluster distance metrics and perform clustering
7. Report - generate popup reports
8. Help - the help menu - local and server based

File menu

This provides commands to select samples for the current, X, Y, X- and Y-sets, and EP-set of samples. It invokes a popup sample chooser. Finds proteins by SwissProt Accession number (SP-ACC) or SwissProt ID (SP-ID). Save and restore the data-mining state. Update ProtPlot and its data from the Web server.

- Select samples ▶
  - Select Current PEP sample - select sample for expression mode
  - Select X PEP sample - select X sample for use in X/Y ratio mode. If sure sample sets mode is on, then define the X set.
  - Select Y PEP sample - select Y sample for use in X/Y ratio mode. If sure sample sets mode is on, then define the Y set.
  - Select Expression List of samples - the subset of samples is used in Expression Profile Plots and in clustering
  - ---
  - List sample assignments - for current, X, Y, and X set, Y set and EP list
  - Assign X set name - to assign a new condition name to the X samples set
  - Assign Y set name - to assign a new condition name to the Y samples set
- Find proteins ▶
  - Find protein by SwissProt Acc - if the SP-ACC is in the database

ProtPlot Data Format



Format of .prp ProtPlot data files

TMAP home | Introduction | Using ProtPlot | Menus | Estimating expression | Download | PEP file format | PDF documents | History of ProtPlot | Latest version

The ProtPlot data is contained in a set of tissue- and histologic-specific .prp files. The set of .prp files constituting the ProtPlot database is included when you download ProtPlot. A .prp file is named using the following convention:

```
[tissue name]_[histological state]_[data_typ].prp
where:
   [tissue name] could be: brain, pancreas, or prostate, etc. For the
      GeneLogic Inc. "DiscoveryBase" DB, a LIST OF tissue names.
   [histologic state] can be a histologic category: normal, premalignant, or cancer.
```

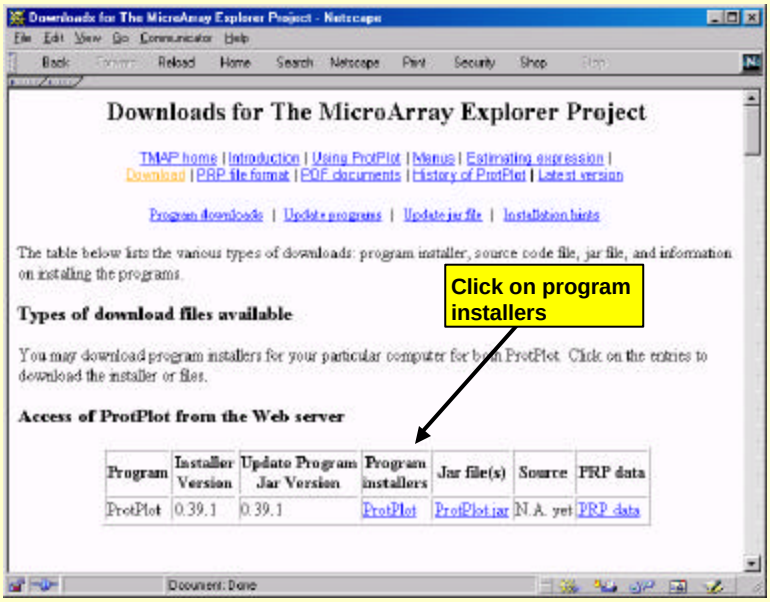
The .prp format has the following tab-delimited format (without the quotes added here for clarity). The first row is the tab-delimited list of field names followed by the tab-delimited corresponding data. The order of the columns is not important. Additional columns may be included in the files, but are ignored if the key words are different from any of the keywords in the following list:

```
'pI'    'Molecular Mass' 'SP-ACC' 'SP-ID'      'MaxESTrope' 'Tissue' 'Family'
6.78    91398            P01265   HUFI_STRAB   0.9840748*   30       1
5.63    99424            P07327   ADC2_STRAB   0.6621861*   30       1
8.39    3929             P42877   HUFI_STRAB   0.9840751*   30       1
```

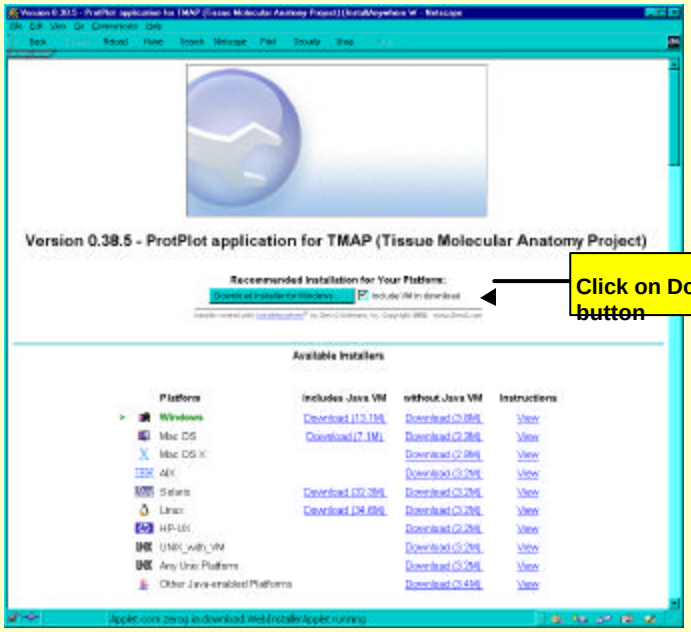
where the following lists the files (case-independent) and their alternate names:

1. 'pI' is the estimated isoelectric point of the protein. It is a decimal number (e.g., 4.61). Alternate names: (pI, pIe)
2. 'Molecular Mass' is the molecular mass in Daltons (see AlloDaltons). Alternate names: (Mw, Molecular Mass)
3. 'SP-ACC' is the SwissProt Accession number. Alternate names: (SP_ACC, SP-ACC, SPACC, 'SwissProt Acc')
4. 'SP-ID' is the SwissProt ID. Alternate names: (SP_ID, SP-ID, SPID, 'SwissProt ID')
5. 'GB-ID' is the GenBank ID (optional field). Alternate names: (GB_ID, GB-ID, GBID, 'GenBank ID')
6. 'MaxESTrope' is the derived expression in the range of 0.0 to 1.0. Missing proteins are not entered. (NOTE is the master protein index computed across all samples, 0.0 indicates there is no protein for a particular tissue when refering to its expression as part of the expression profile.) See discussion on how MaxESTrope is computed for ProtPlot. Alternate names: (GB_ID, GB-ID, GBIDMaxESTrope, 'Max EST rope', estRope, espe, espression)
7. 'Tissue' specifies the tissue(s) that constitute the samples. Alternate names: (Tissue, 'Tissue Name'). It is either:
   1. a tissue number from the 'tissueNamesFile.txt' file, or
   2. a hexadecimal bit pattern of several tissue numbers (eg. a mixture of tissues), then it is represented by the sum of (2**tissue(i)) for a set of n tissues
8. 'Family' is the protein families that the protein belongs to. It is a hexadecimal bit pattern of several protein family numbers from the 'familyNamesFile.txt' file. The number of families is sum of (2**family(i)) for a set of n families. Alternate names: (Family, 'Family Name', 'Protein Family', 'Protein Family Name'). [FUTURE] (This data is optional).

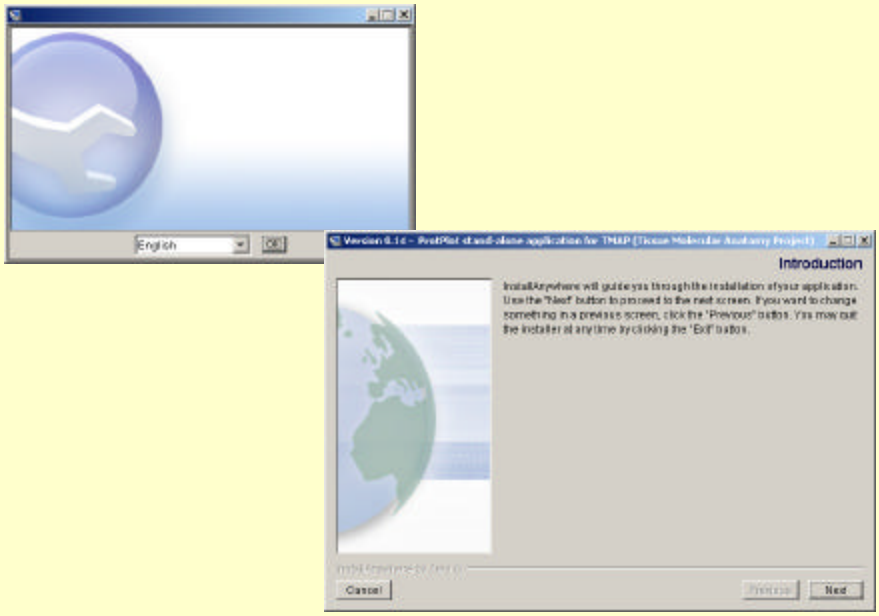
Download ProtPlot



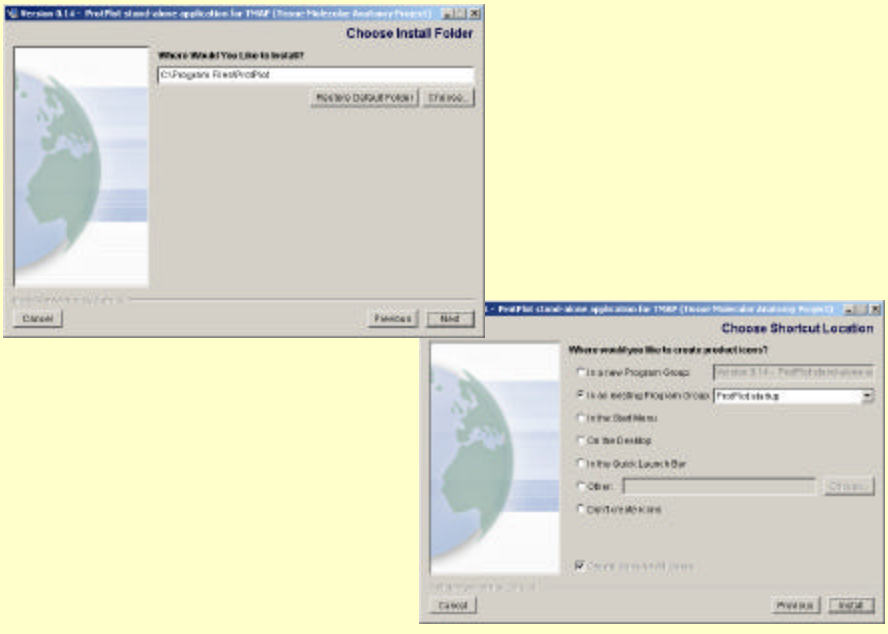
Download ProtPlot Installer



Installing ProtPlot



Installing ProtPlot (continued)

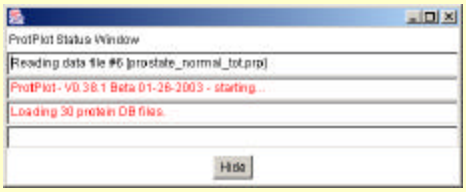


Finished Downloading ProtPlot

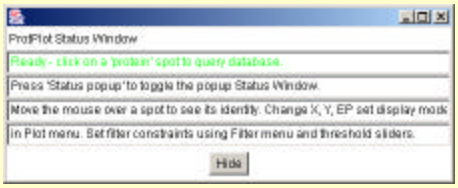
Starting ProtPlot - Click on the Startup Icon or Use the Start menu

A) Click on ProtPlot Startup icon

B) Displays the loading status



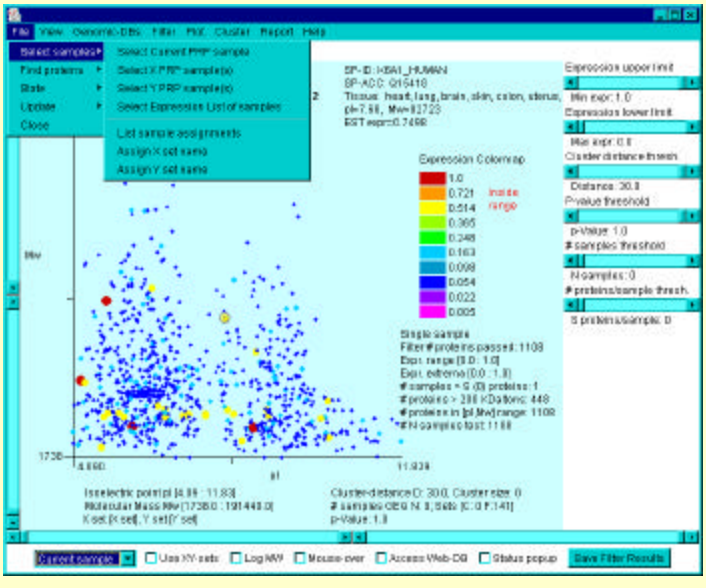
C) Press Hide button to remove



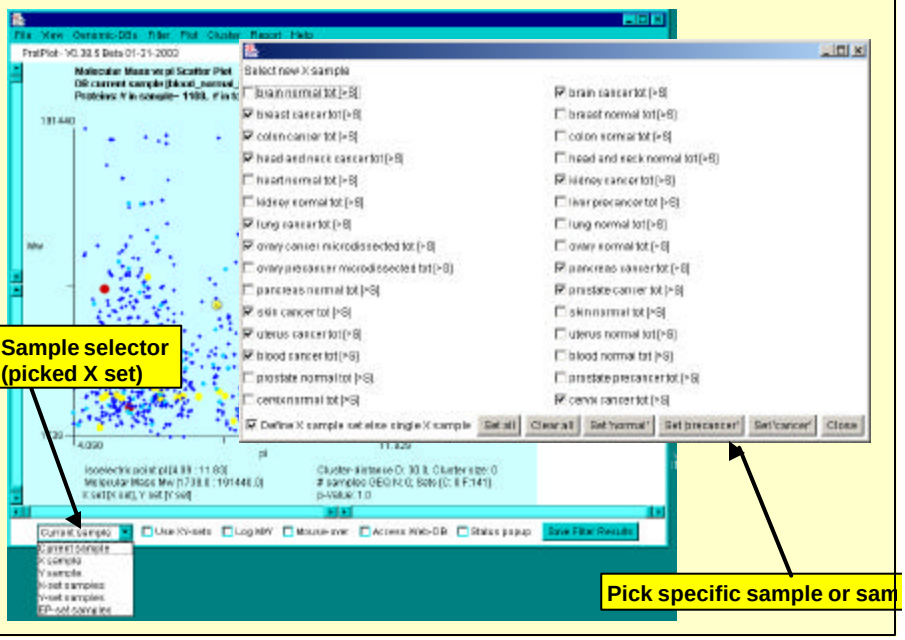
ProtPlot Menus

- **File** - select samples, save the state and quit
- **View** - select viewing options
- **Genomic-DBs** - enable access to popup Web genomic databases
- **Filter** - select protein data filter options
- **Plot** - select primary data mining and scatterplot display options
- **Cluster** - select cluster distance metrics and perform clustering
- **Report** - generate popup reports
- **Help** - popup help menu

File Menu - Selecting Single Samples, the X-set, Y-set or EP-set of Samples



File Menu - Selecting Samples using Choice Menu



Selecting Subsets of Samples for Experiments

- **Current Sample** - to look at the expression for any individual sample.
  E.g., *prostate_cancer*

- **Sample X and Sample Y** - to look at the ratio of *exprX/exprY* where the protein for which the ratio is defined has expression in both the X and Y individual samples.
  E.g., X is *prostate cancer* and Y is *prostate_normal*

- **X set of samples and Yset of samples** - to look at the ratio of *Mean-exprX / Mean-exprY* where the protein for which the ratio is defined has expression in both the X and Y samples for at least 1 sample in X and at least 1 in Y.
  E.g., X set is all *cancer* and Y is all *normal*

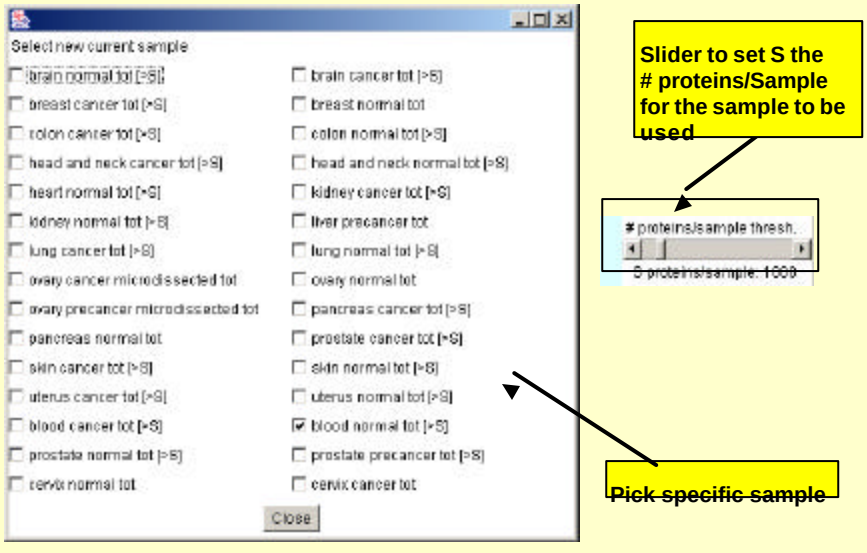
- **Expression Profile set of samples** - to look at the expression profile (EP plot or EP report) for any protein. The scatter plot shows mean EP expression.
  E.g., EP is all samples, or EP is all *cancer, etc.*

Plot Display Mode Rules

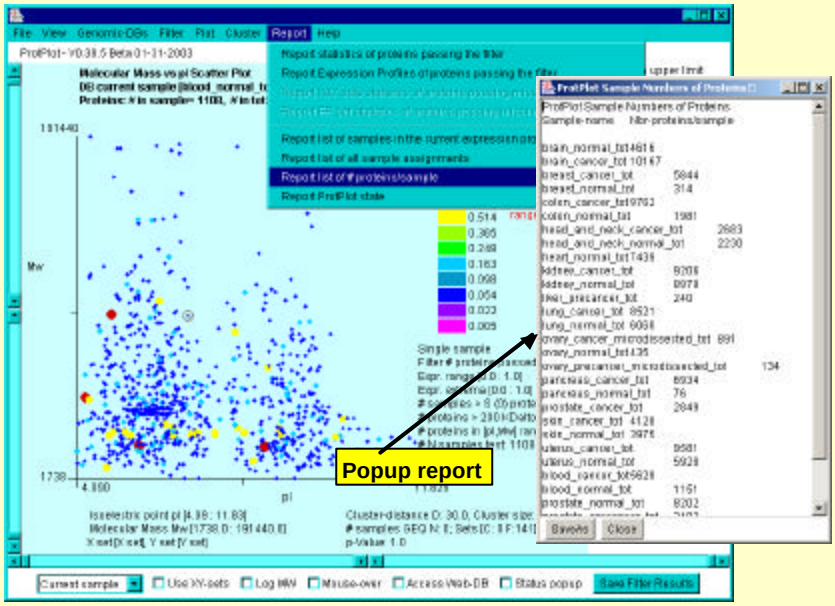
All proteins in the Master Protein Index (mPid) are displayed except for the following:

- In single sample or EP expression mode, do not show missing proteins
- In X/Y sample mode, do not show proteins that are missing in X but present in Y or vice versa. However, if the View option to display this missing data is enabled, then show the missing data as gray spots.
- In X-set/Y-set samples mode, do not show proteins unless they meet the sizing criteria N for both X and Y if enable or if using the missing sets > N filter.
- Normally, plot proteins in a (**Mw vs. pI**) scatterplot
- If in one of the X/Y ratio modes, may plot (**X vs. Y**) expression scatterplot instead of (Mw vs. pI)

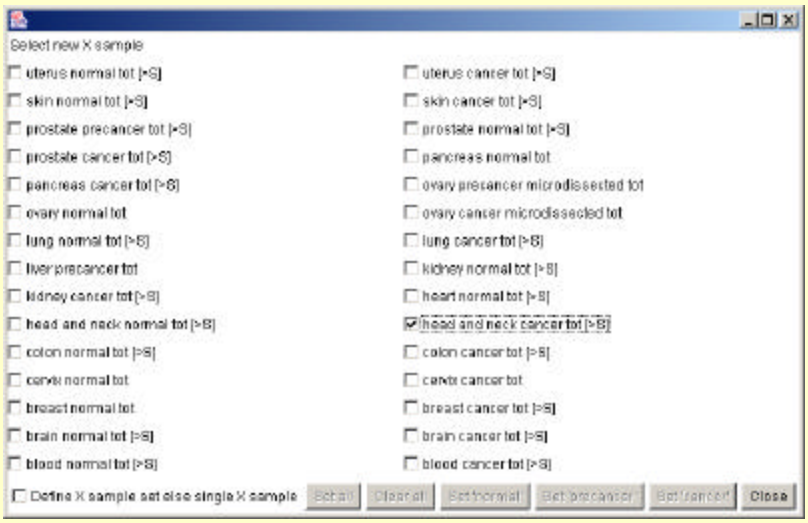
Selecting the Current Sample (those with [>S] have more than S proteins/sample)



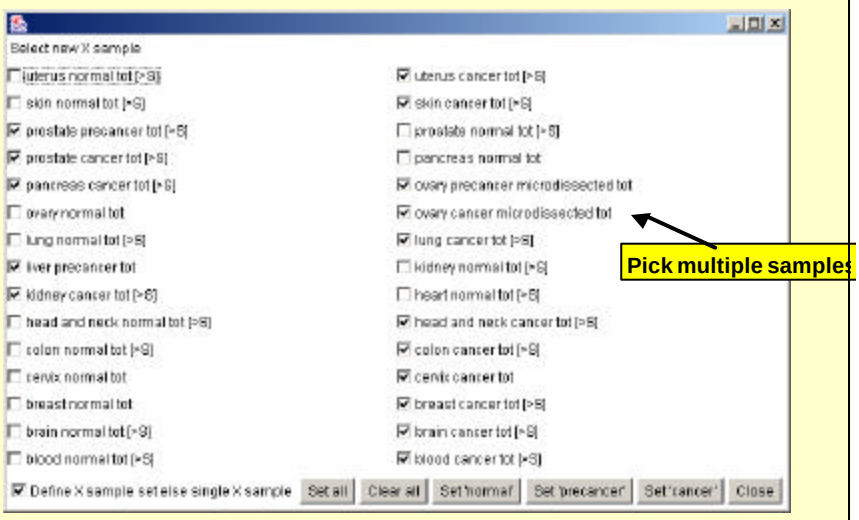
Report Menu - Listing # Proteins in All Samples



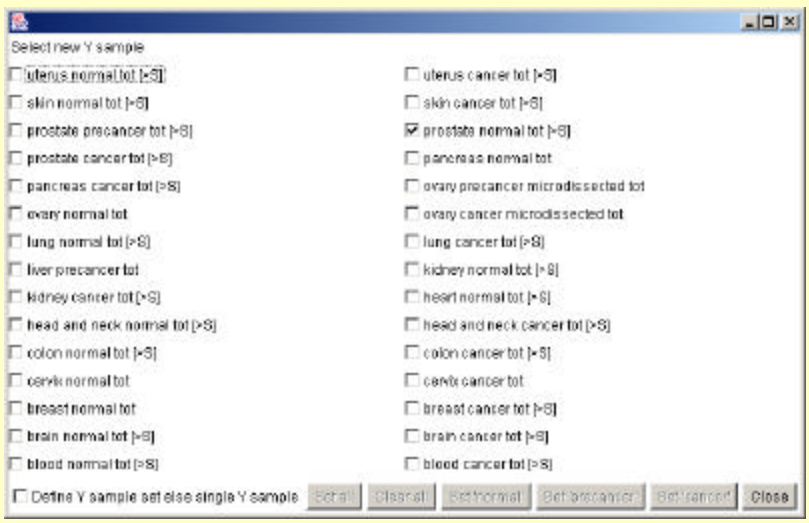
Selecting the X Sample



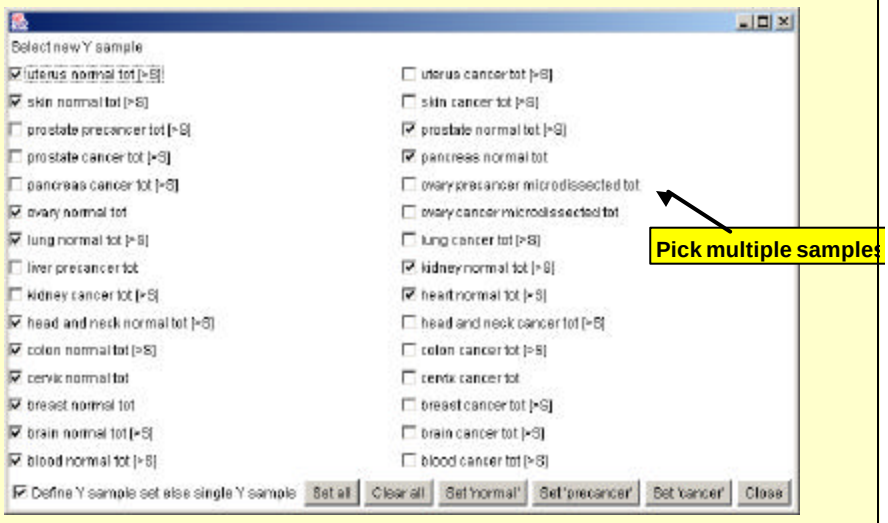
Selecting the X-set of Samples



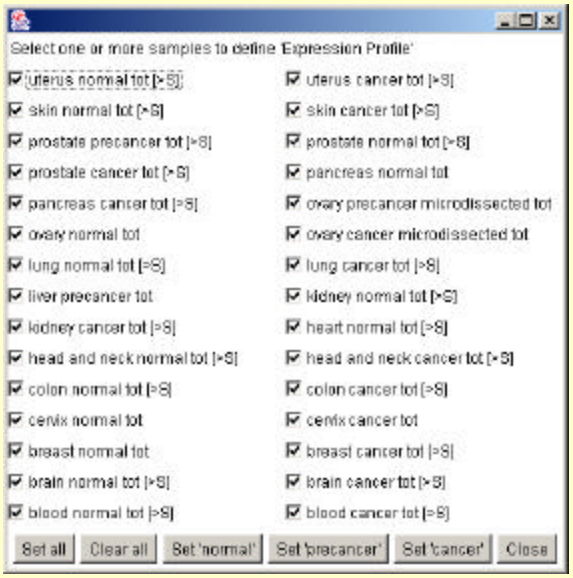
Selecting the Y Sample



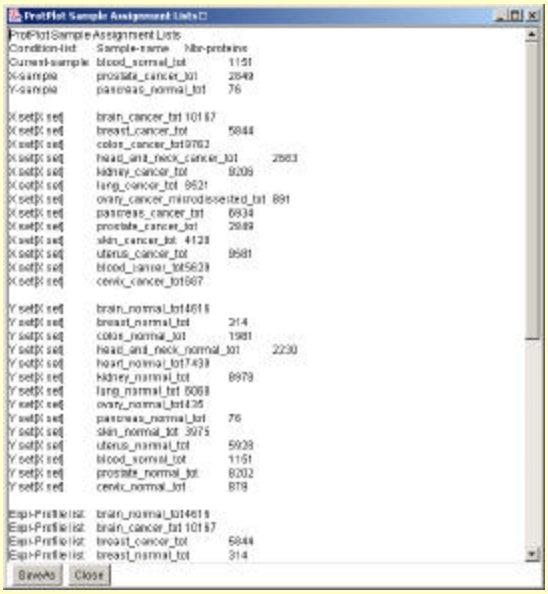
Selecting the Y-set of Samples



Selecting the Expression Profile (EP) Set of Samples

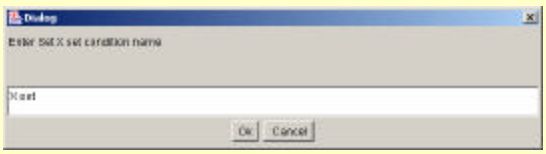


Listing Sample Assignments

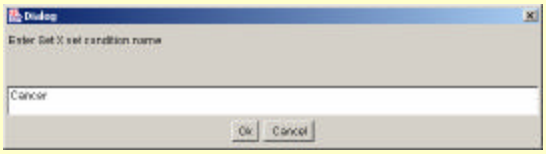


Defining the X and Y Condition Set Names

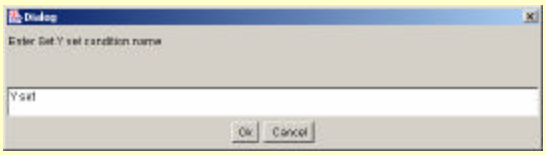
A.1
(default X set)



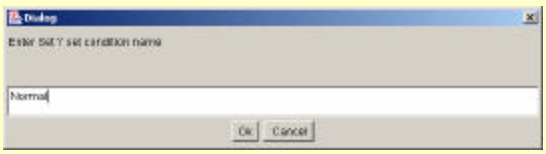
A.2
set to 'cancer'



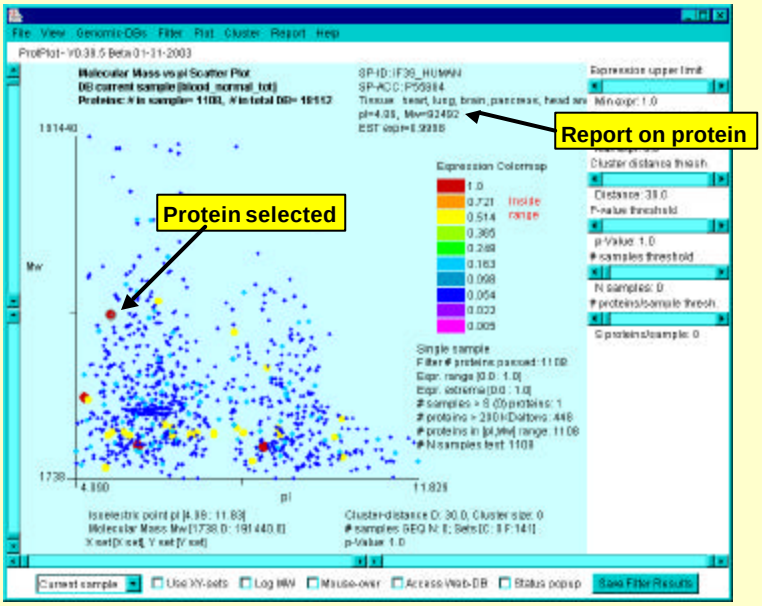
B.1
(default Y set)



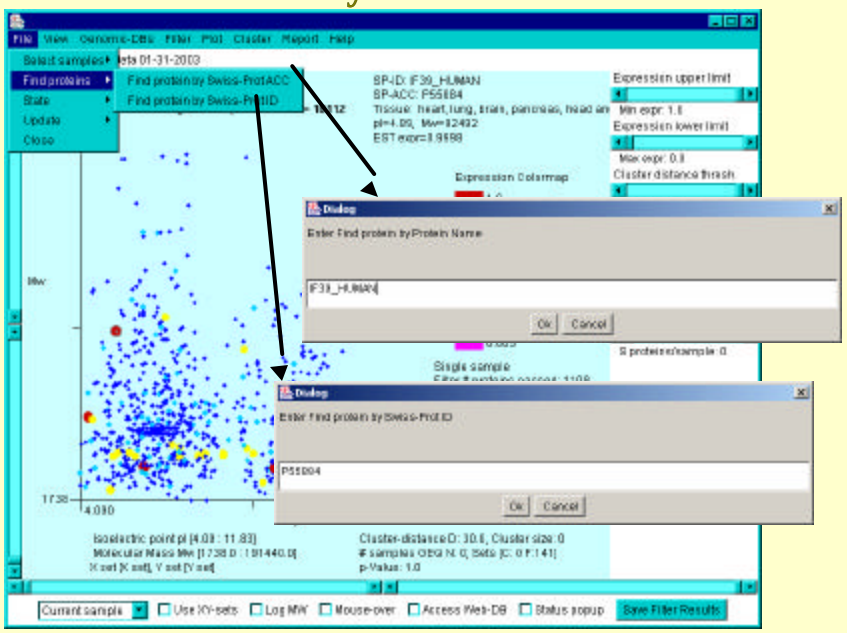
B.2
set to 'normal'



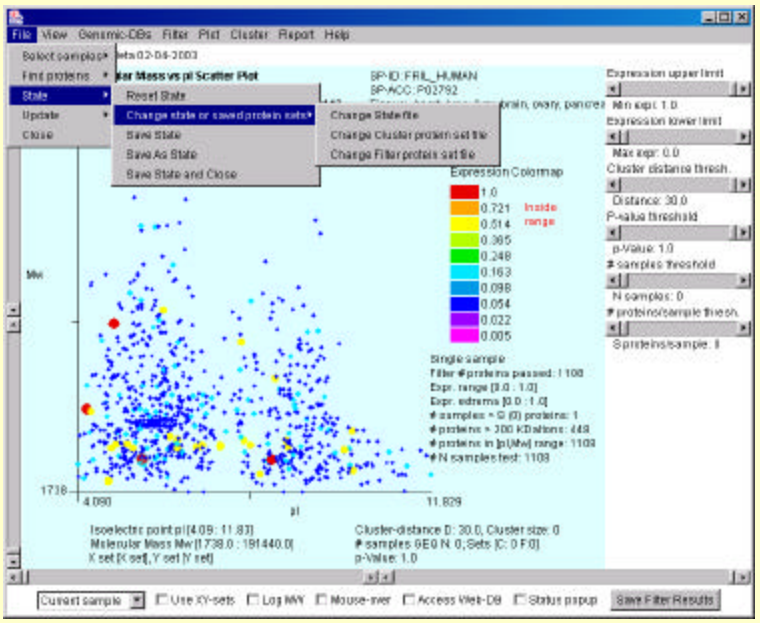
Click on a Spot to Select the Protein



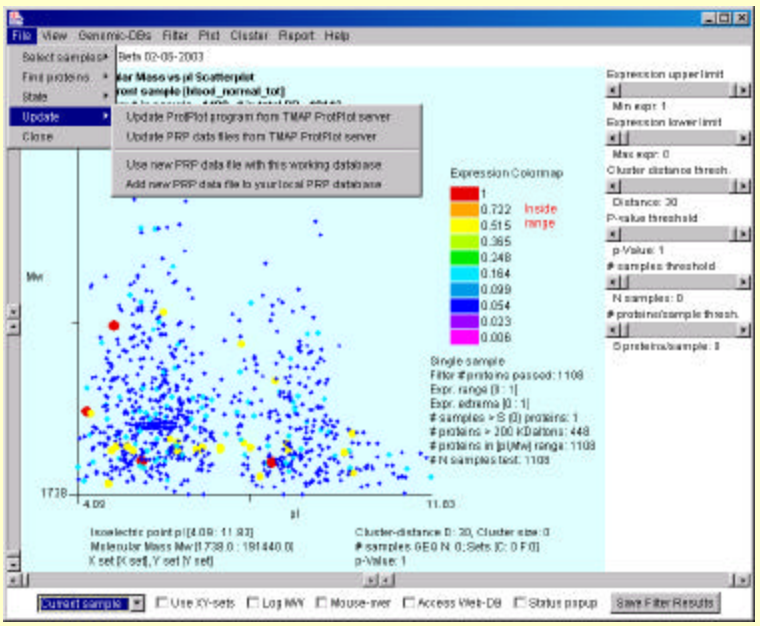
Select a Protein by SwissProt ID or ACC



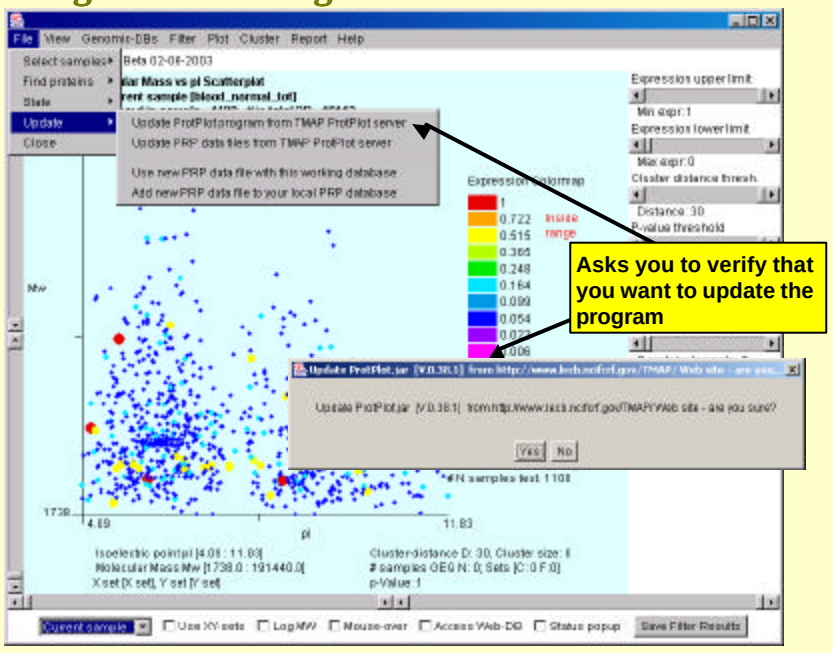
File Menu - Save & Restore the Data Mining State



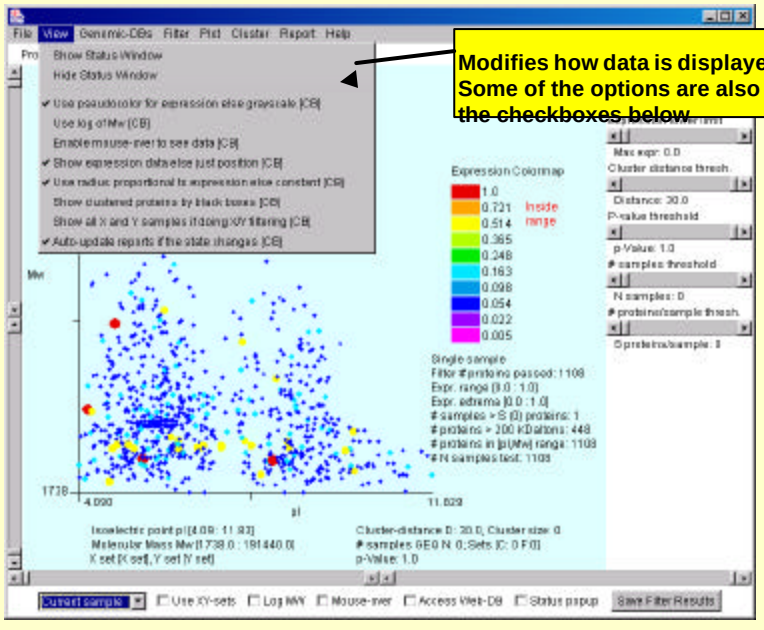
File Menu - Updating the Program and PRP data



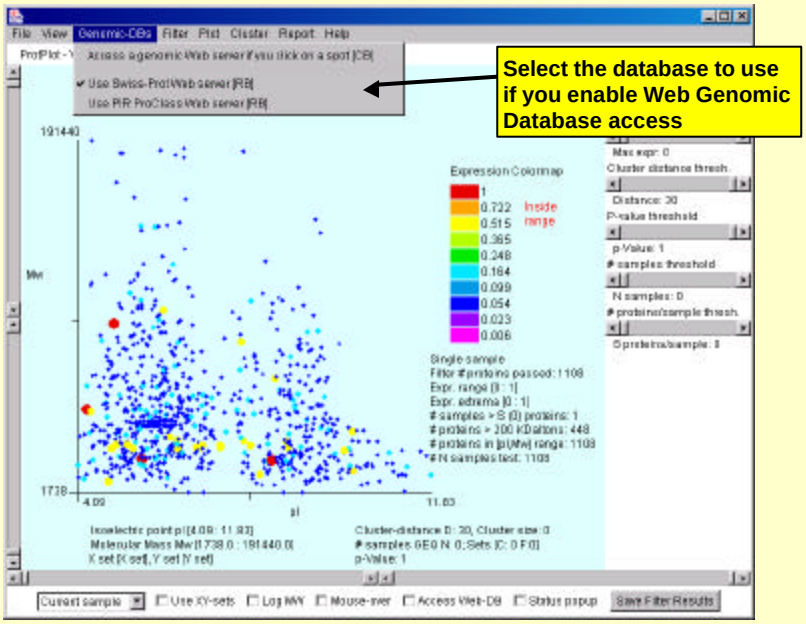
Updating ProtPlot Program from the Proteom Server



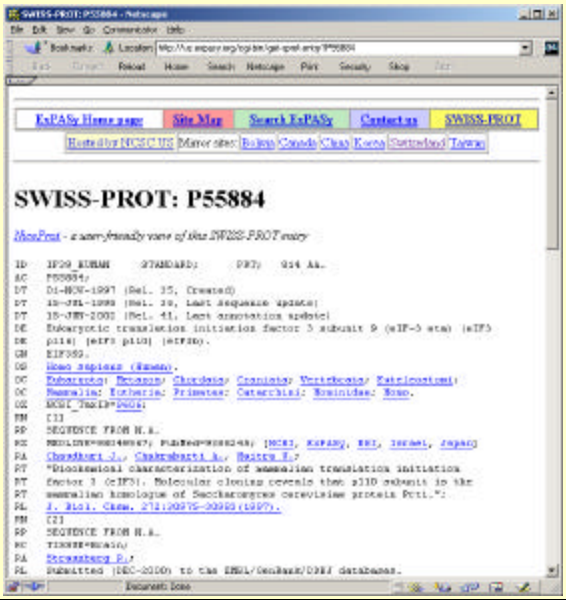
View Menu - Display Options



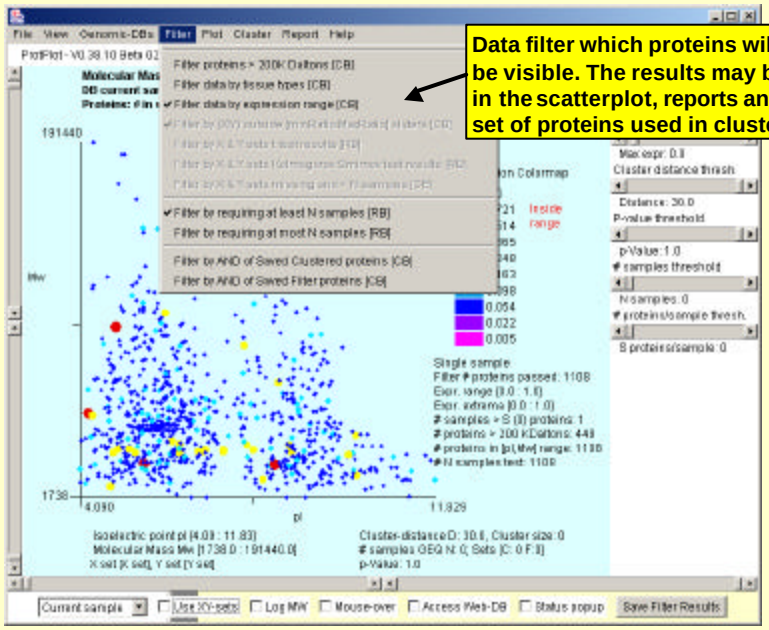
Genomic-DBs Menu



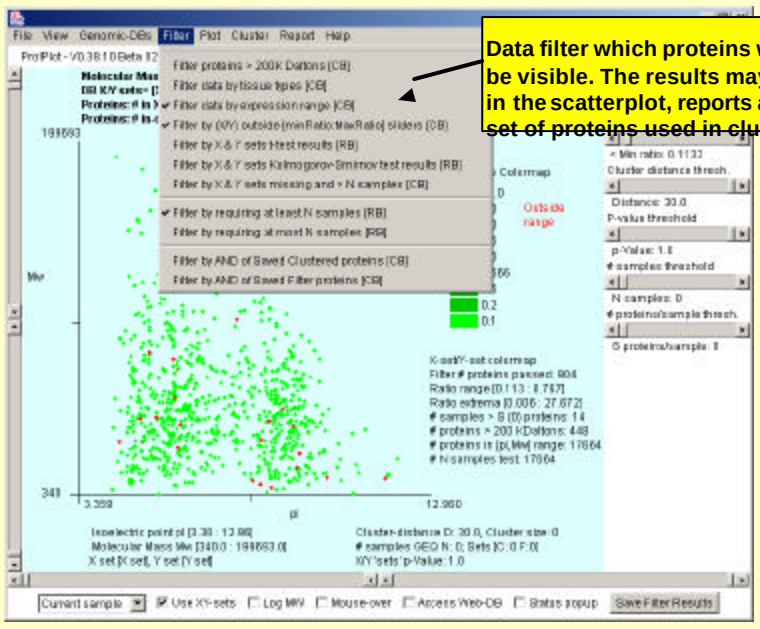
Bringing up a Genomic Server by Clicking on Spot if you Enabled Genomic DB Access



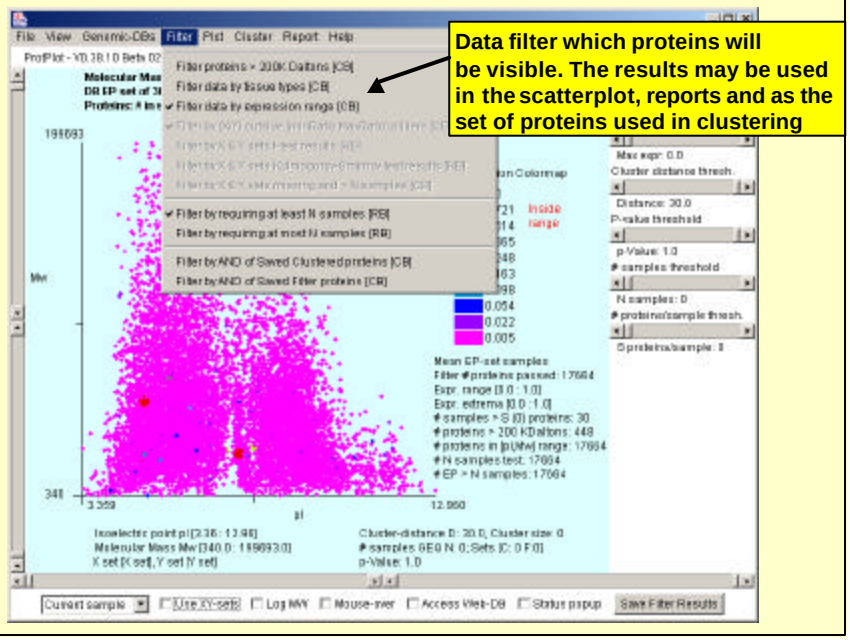
Filter Menu - Data Filter Options for Single Sample



Filter Menu - Data Filter Options for X/Y Ratio



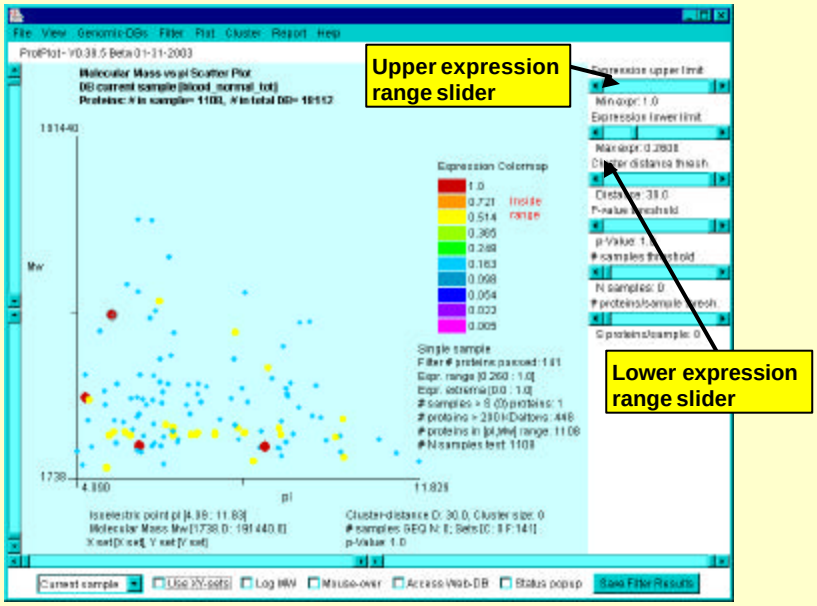
Filter Menu - Data Filter Options for EP-set



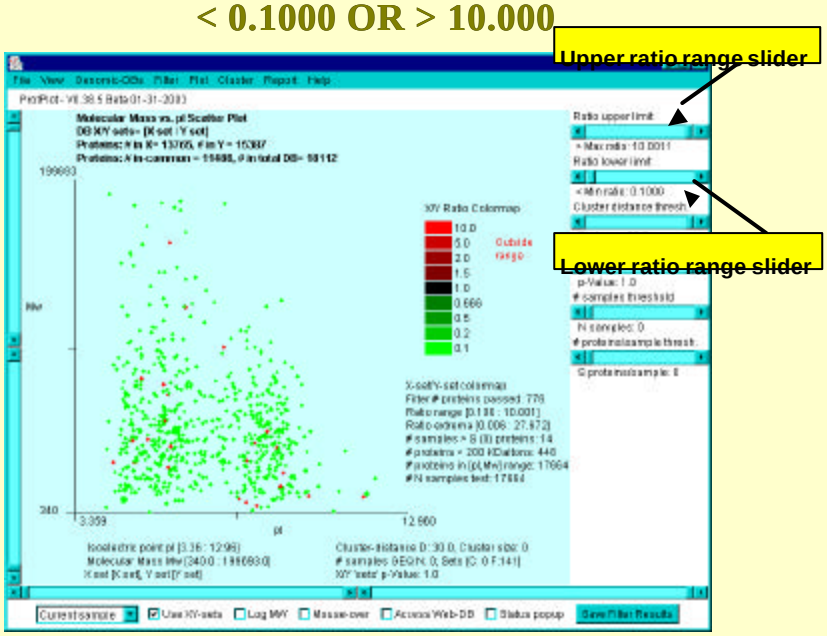
Filter Types - Available

- **By Proteins > 200 Kdaltons , Mw and pI within ranges**
- **By tissue types**
- **By expression value range**
- **By expression X/Y ratio range (either inside or outside range)**
- **By t-Test of X-set and Y-Set samples < p-value threshold**
- **By min # samples in X &Y or EP sets > N samples threshold**
- **By missing proteins in X or Y set with other set > N samples threshold**
- **By number of samples for the protein > N samples threshold or < N samples threshold**

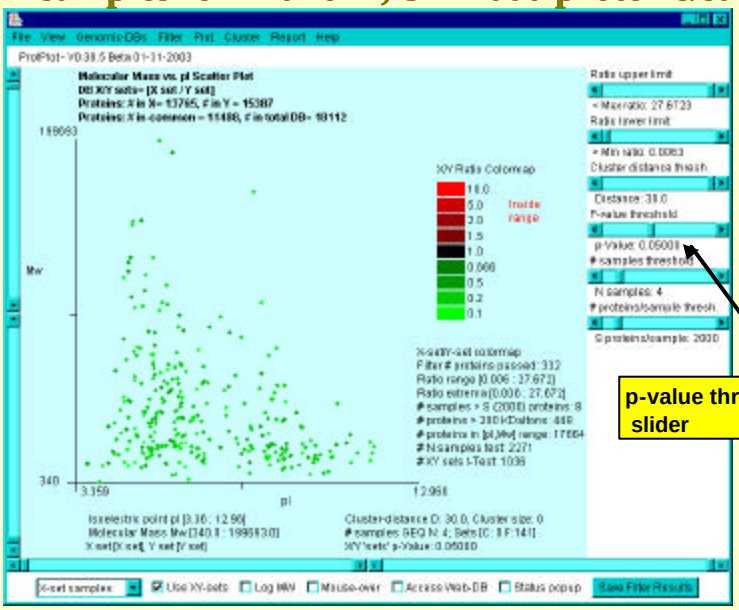
Applying Expression Range Filter [0.455 : 1.0]



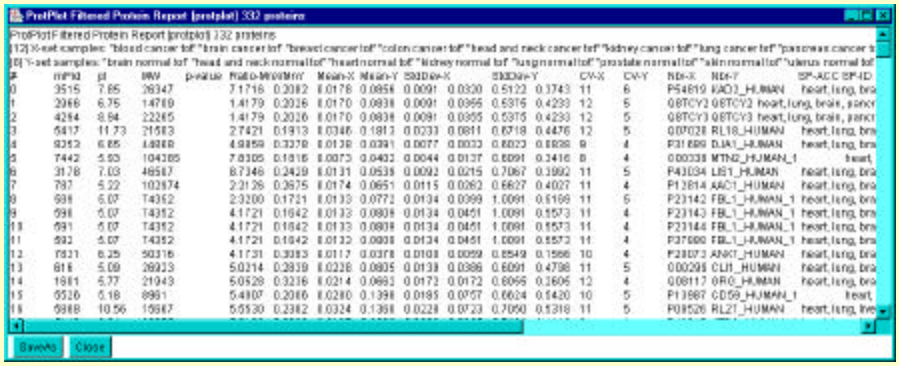
Applying the 'outside' X/Y Ratio Range Filter < 0.1000 OR > 10.000



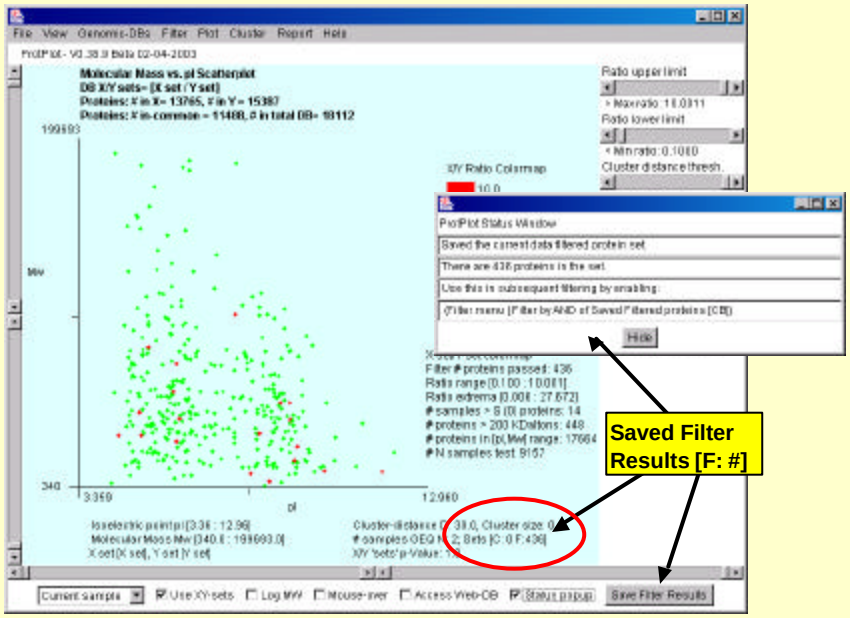
Applying the t-test (p=0.05) Filter X/Y sets
Min 4 samples for X and Y, S>=2000 proteins/sample



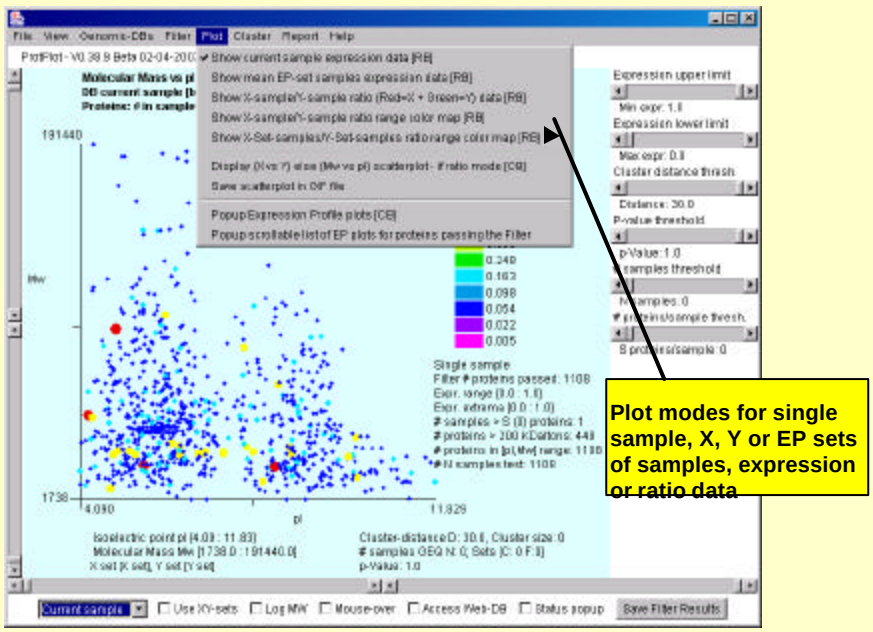
Applying the t-test (p=0.05) Filter X/Y sets
Min 7 samples for X and Y, S>=2000 proteins/sample



Saving Filter Set of Proteins - For Future Filtering



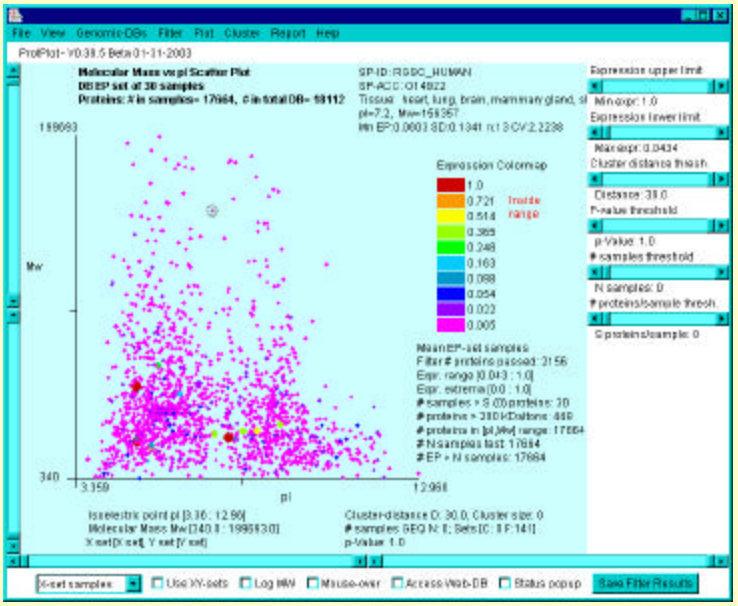
Plot Menu - Display Mode and Options



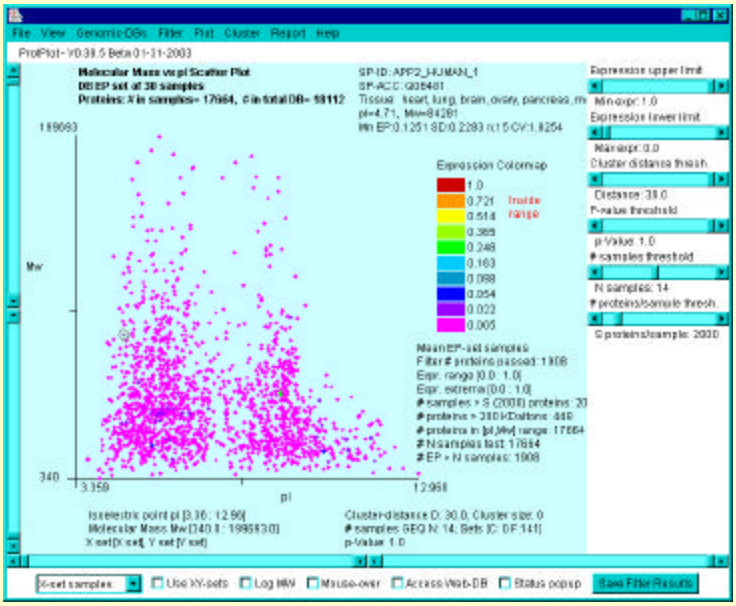
Plotting Display Modes

- **Show Current Sample** - to look at the **expression** for a single sample
- **Show Mean Expression-Profile set of samples** - to look at the **mean expression** for a subset of samples
- **Show X-Sample /Y-Sample Y** - to look at the **ratio** of two individual samples
- **Show X-set samples / Y-set samples** - to look at the **ratio** of *Mean-exprX / Mean-exprY* for two **sets** of samples (X and Y sets)
- If in one of the X/Y ratio modes, may plot (**X vs Y**) expression scatterplot instead of default (**Mw vs. pI**) scatterplot

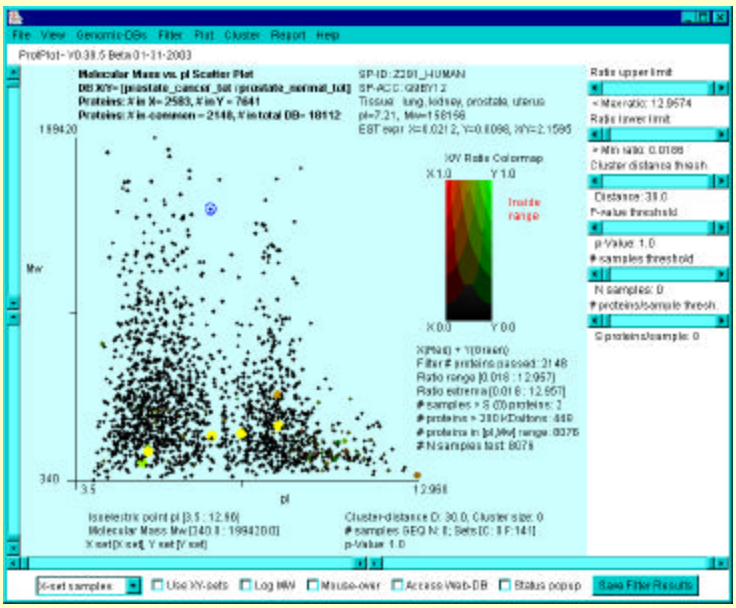
Plot Display Mode - Current Sample



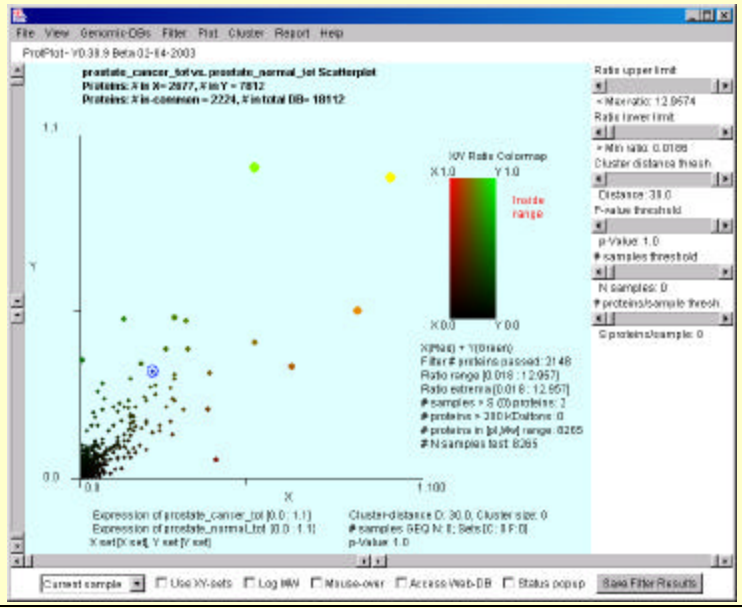
Plot Display Mode - Mean of EP Set of Samples (N >= 14, S >= 2000)



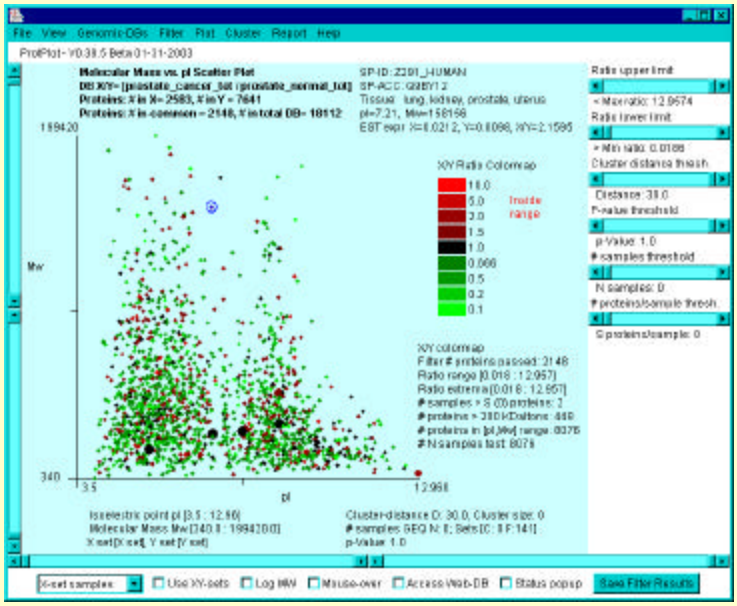
Plot Display Mode - X Sample (Red) + Y Sample (Green)



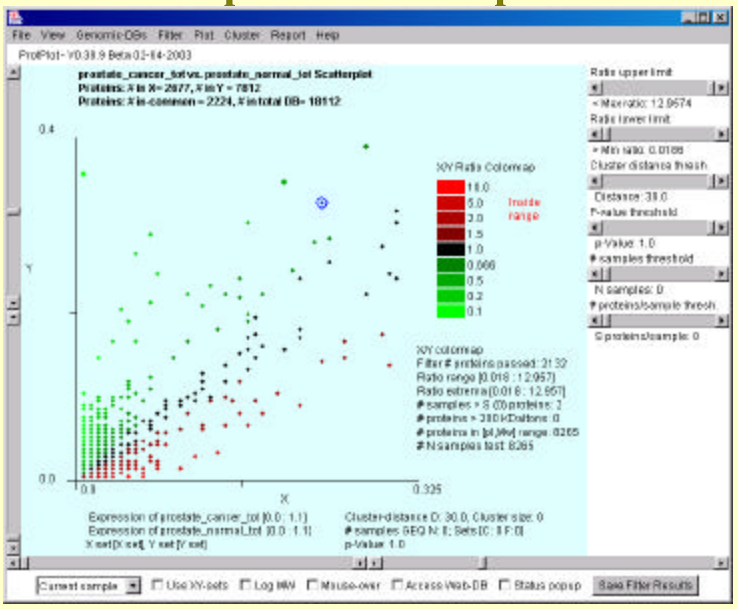
Plot Mode - Sample Xvs Sample Y Expression Scatterplot



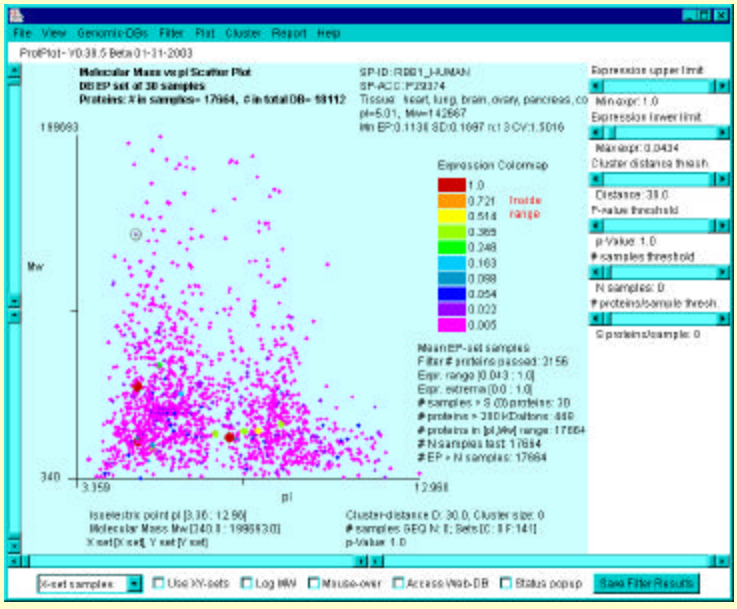
Plot Mode - X Sample / Y Sample Colormap



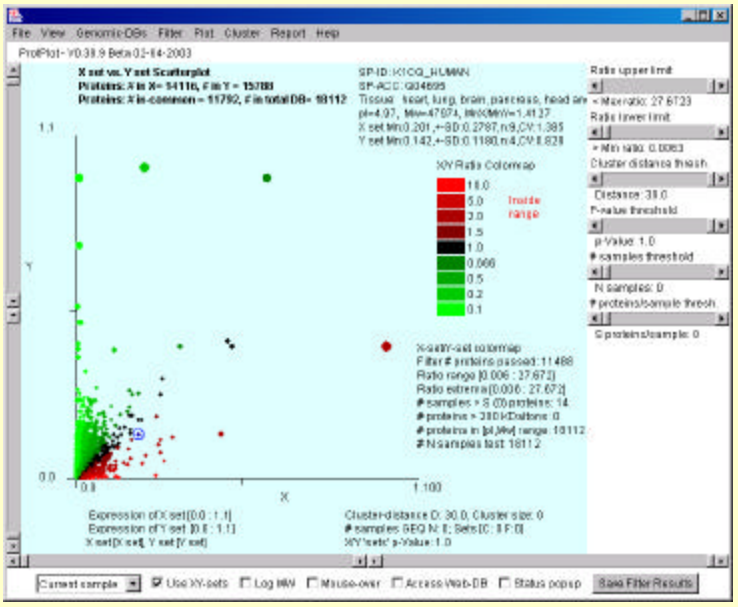
Plot Mode - Sample X vs Sample Y Expression Scatterplot



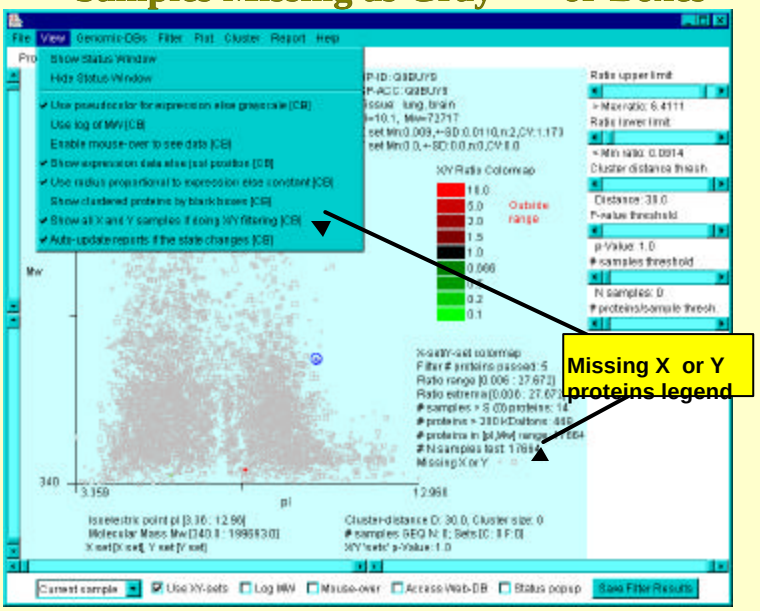
Plot Mode - Mean X-set / Mean Y-set Samples



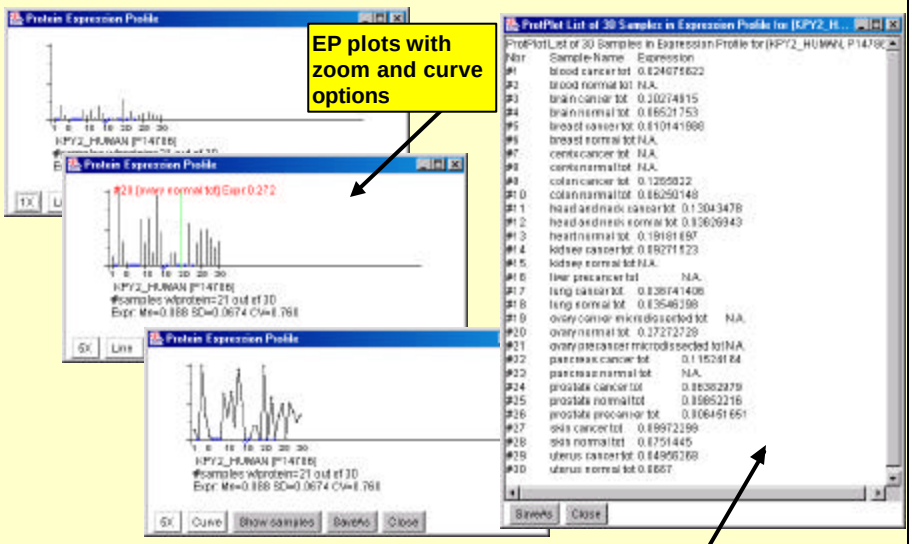
Plot Mode - Mean X-set vs Mean Y-set Expression Scatterplot



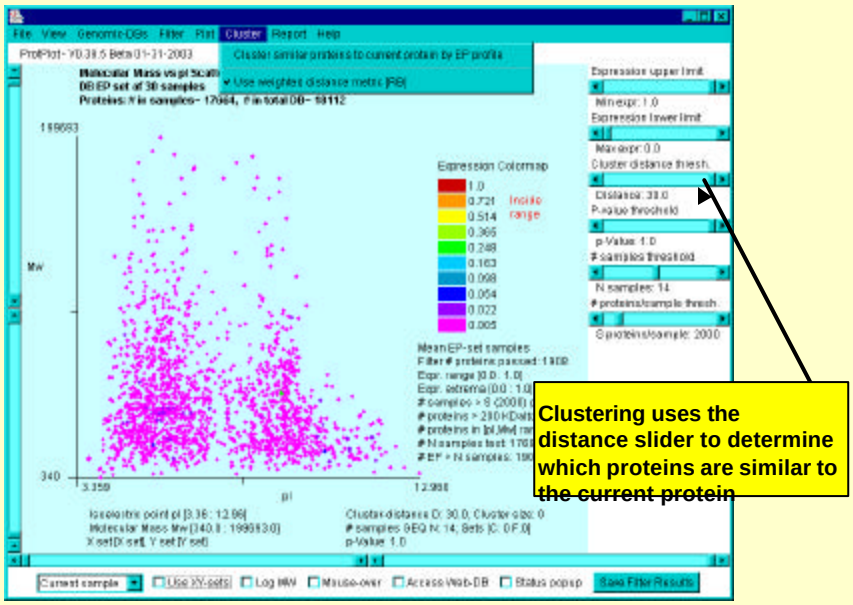
Plot Mode - Showing Proteins With Either X or Y Samples Missing as Gray '+' or Boxes



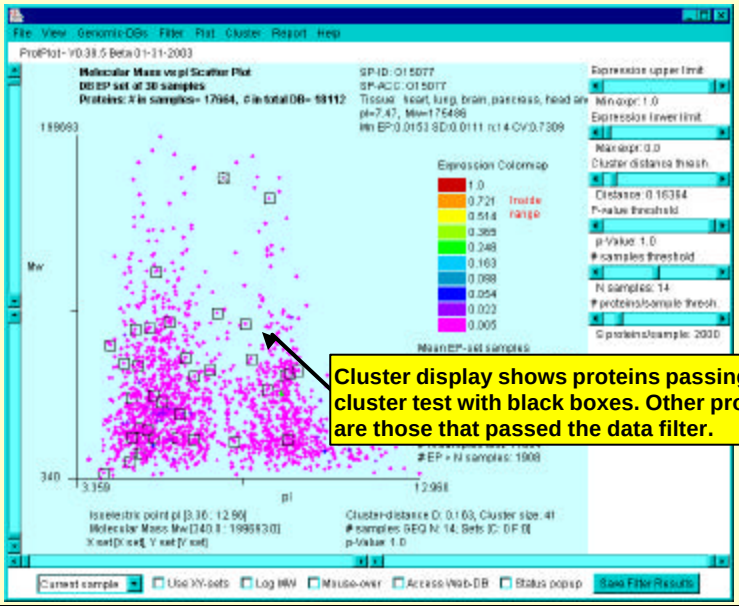
Plot Mode - Popup Expression Profile Plot for 1 Protein - Click on a Different Spot to Change the Plot



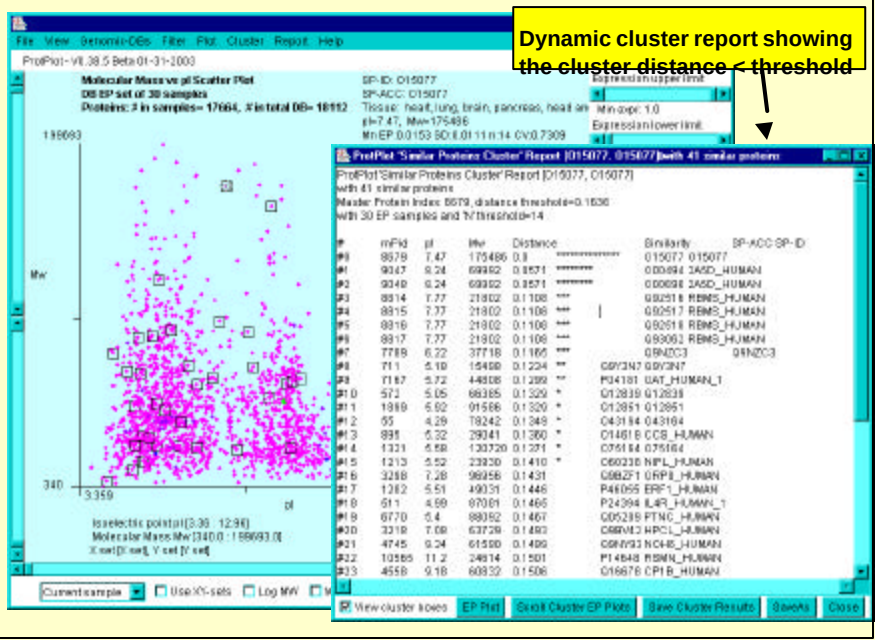
Cluster Menu - Find Proteins with Similar Expression



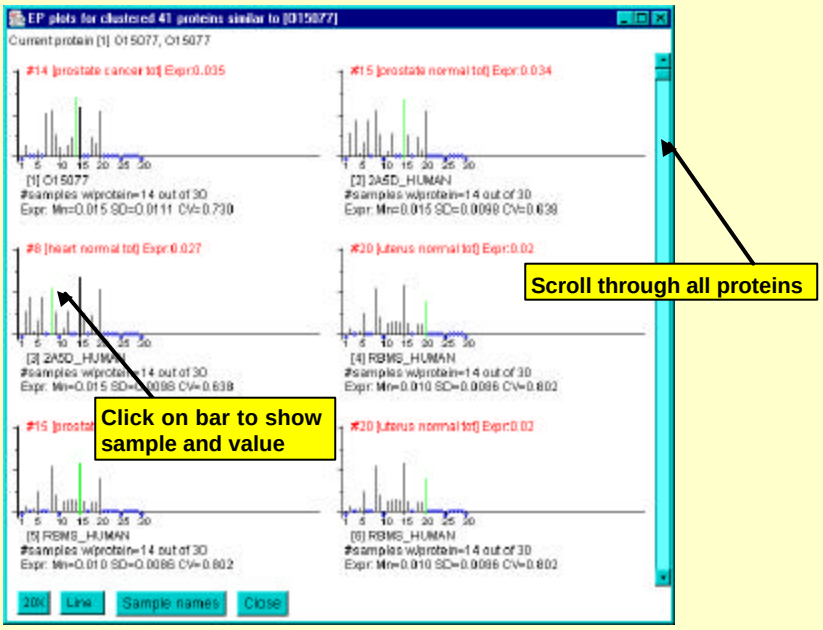
Clustering on Selected Protein - Scatterplot with Cluster Member Proteins Shown with Black Boxes



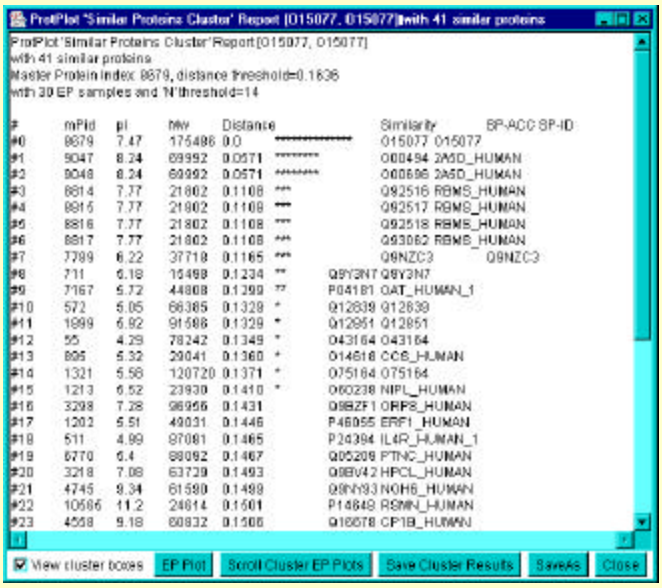
Clustering on Selected Protein (All Samples) D<0.69



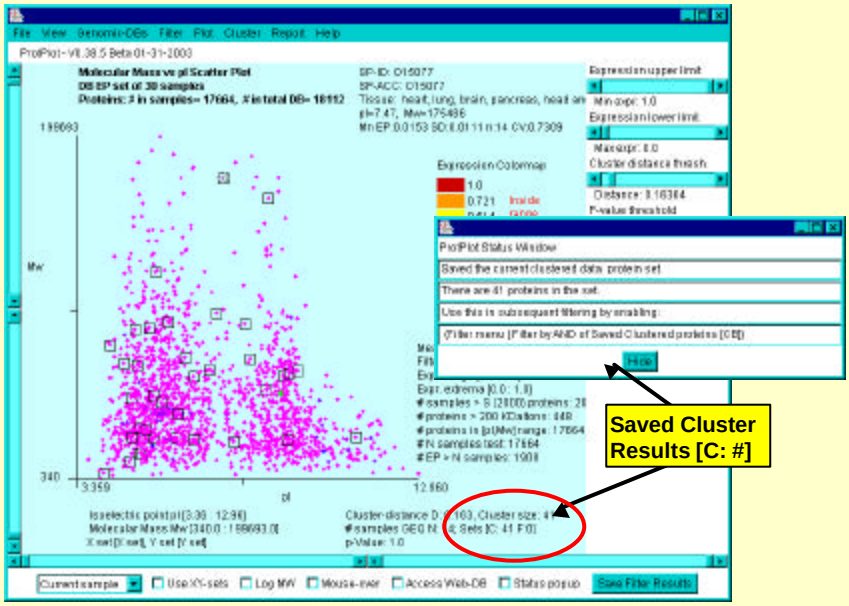
Scrollable EP Plots for Clustered Proteins



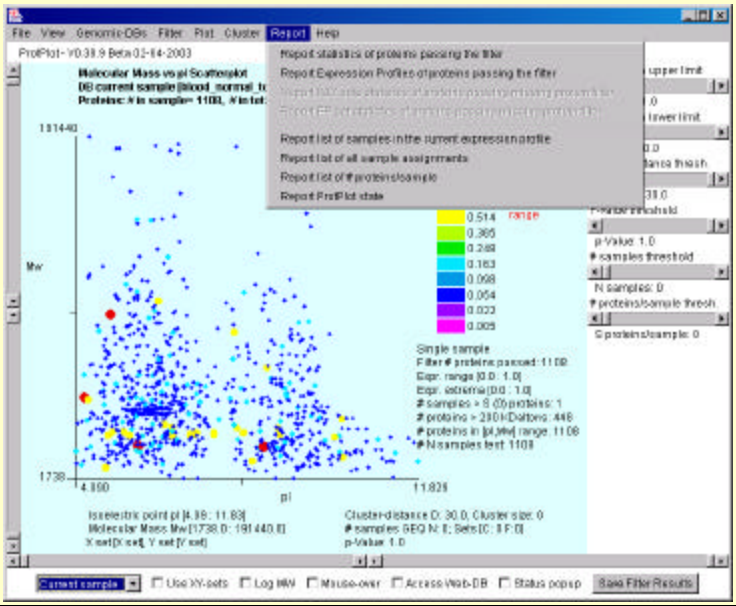
Clustering on Selected Protein - Cluster Report with Silhouette Plot Sorted by Cluster Distance



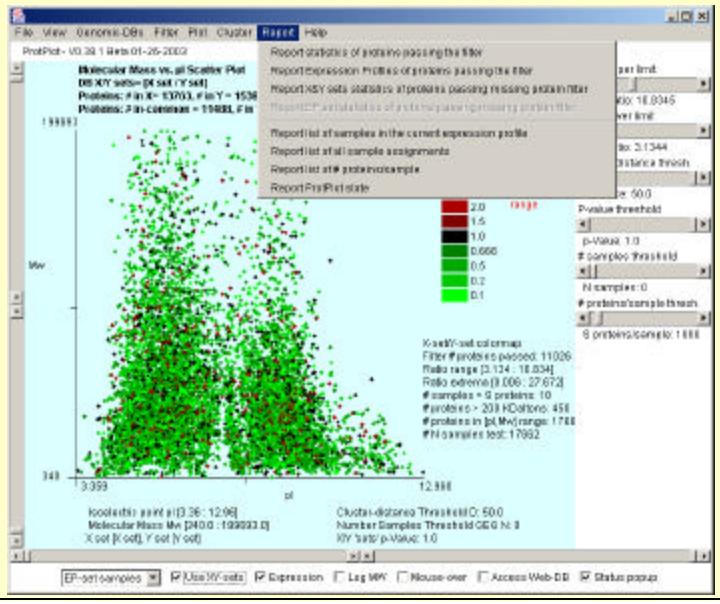
Saving Cluster Set of Proteins - For Future Filtering



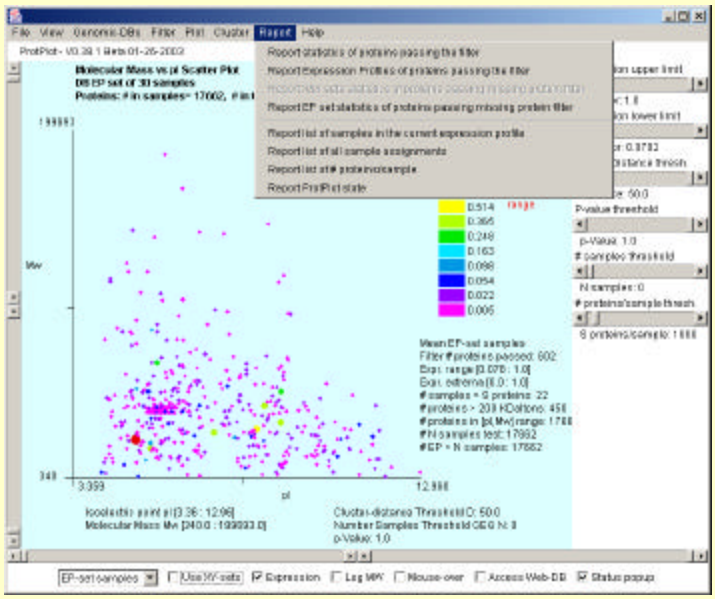
Report Menu - Options are Display Mode Dependent Current Sample Mode



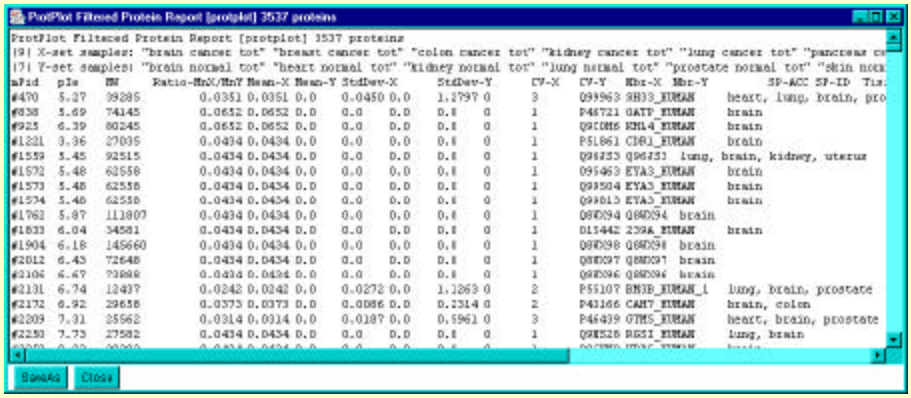
Report Menu - Options are Display Mode Dependent Ratio Mode options



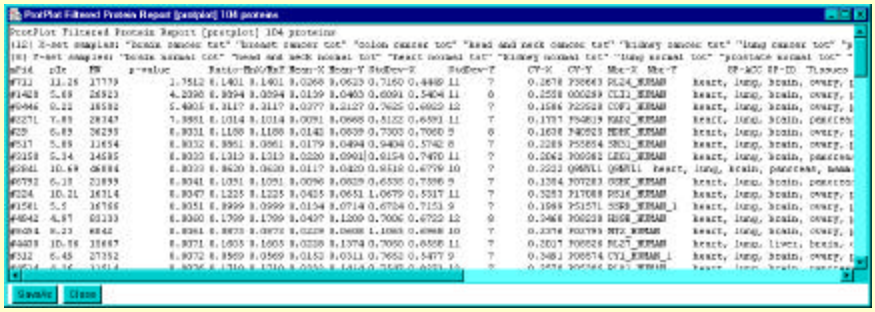
Report Menu - Options are Display Mode Dependent Mean EP Expression Mode



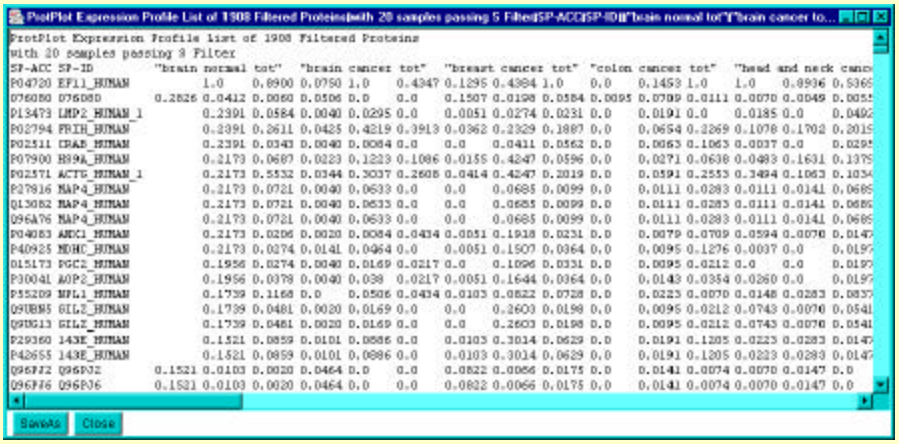
Popup Report for the Filter X/Y sets
Minimum S>=3465 Proteins/Sample



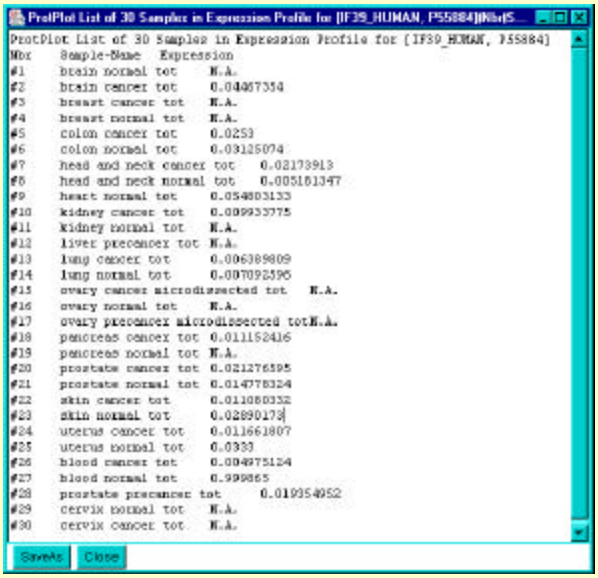
Popup Report for the t-test (p=0.05) Filter X/Y sets
Min 7 Samples for X and Y, S>=2000 proteins/sample



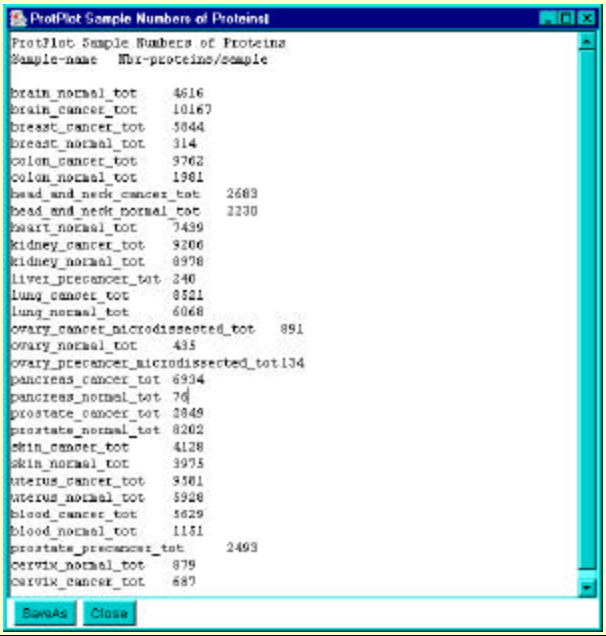
Popup Report Expression Profile Values for Filtered Proteins (min N>= 14 samples, S >=2000)



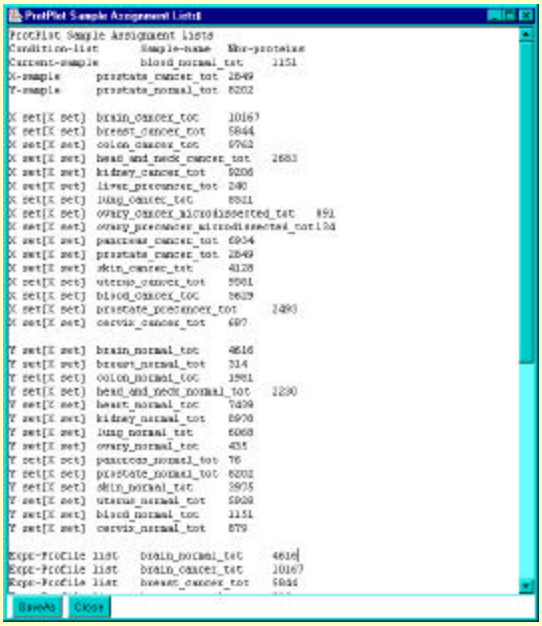
Popup Report of Samples in Expression Profile Set for the Currently Selected Protein



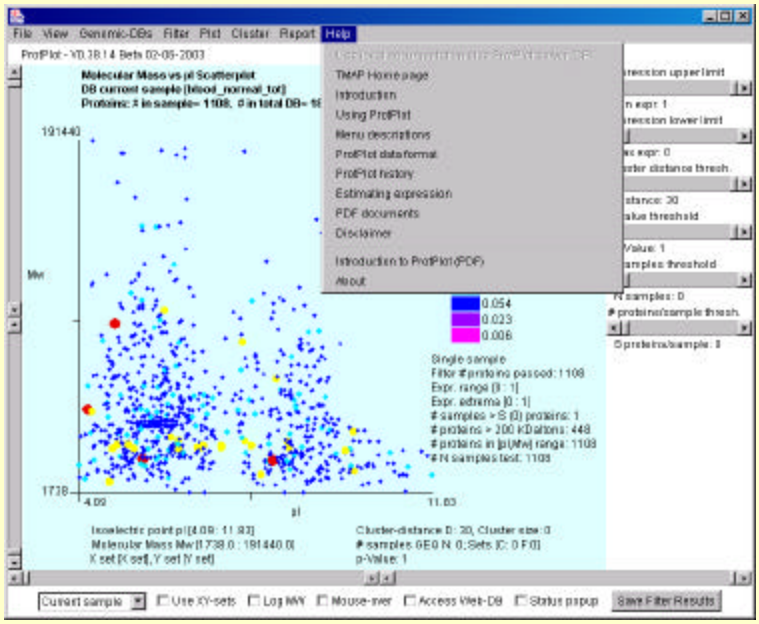
Popup Report # of proteins/sample for All Samples



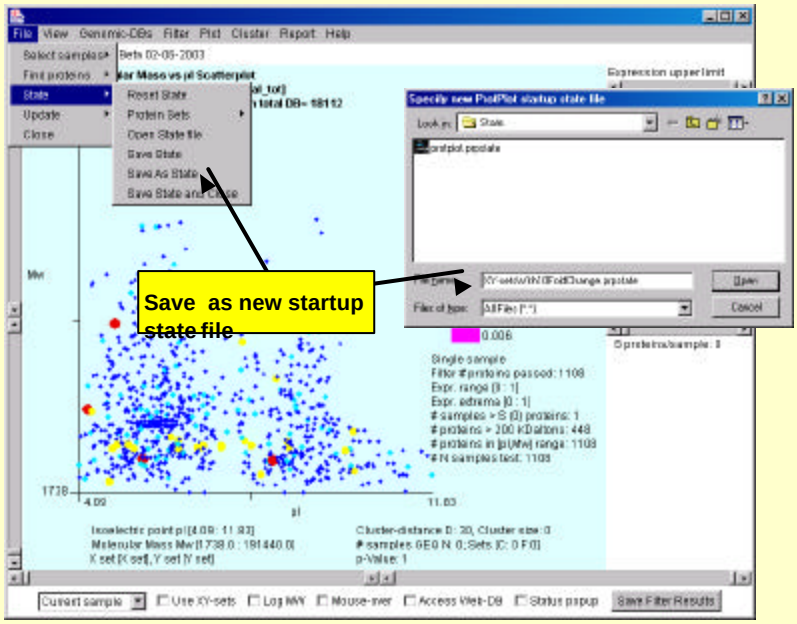
Popup Report All X, Y, EP Sets Sample Assignments



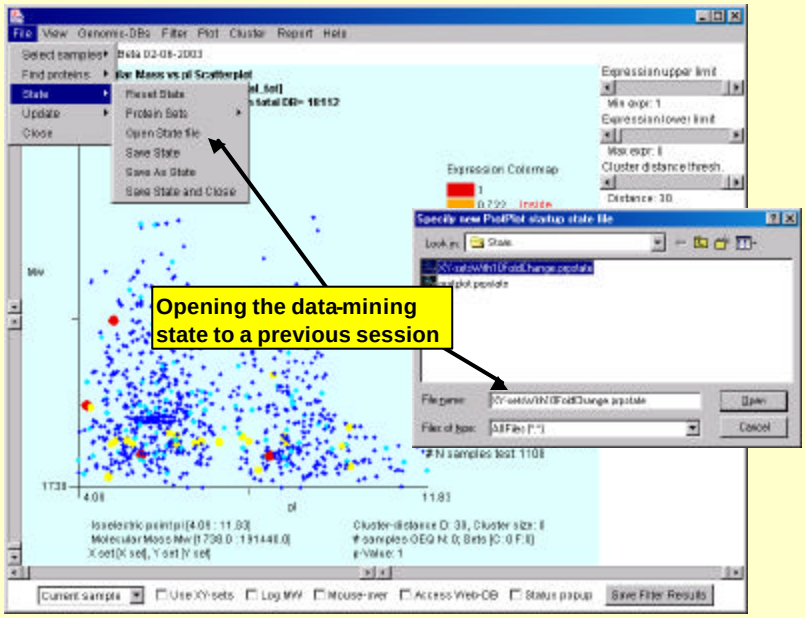
Help Menu - Popup Web Browser Documents



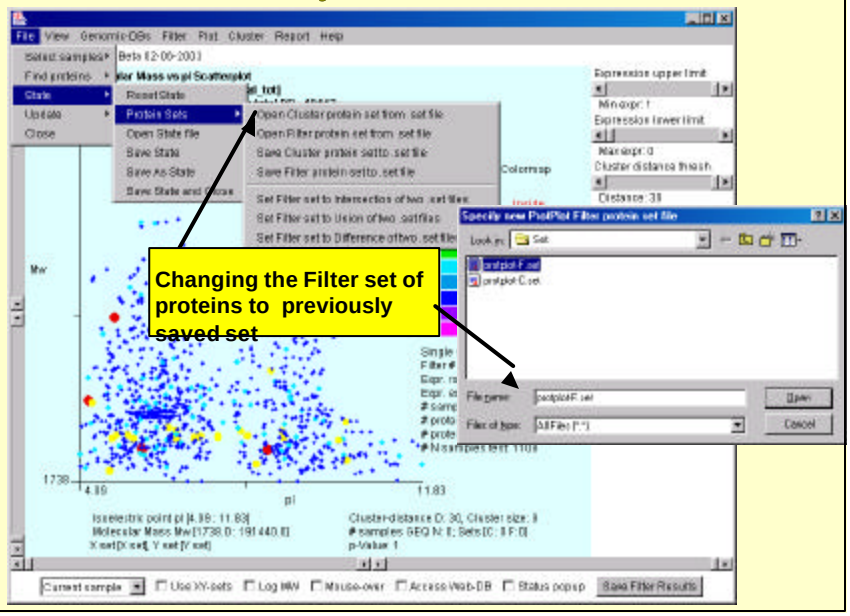
Saving the Current Data-mining Session State



Changing the State to a Previous Data-mining Session



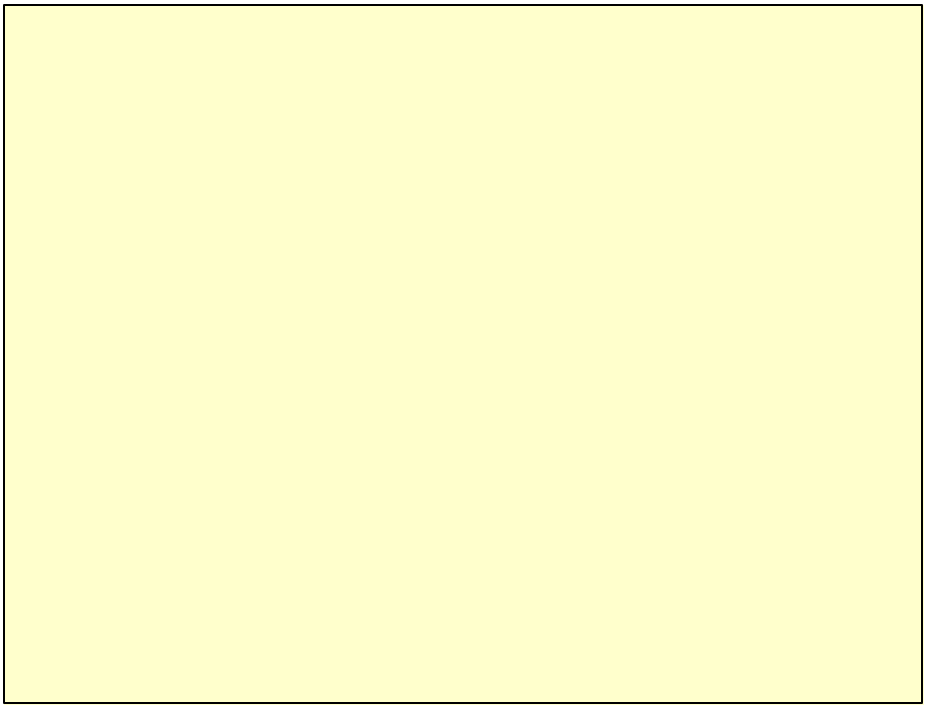
Changing the Filter Set of Proteins to a Previously Saved Filter Set



References

- Medjahed D, Luke BT, Tontesh TS, Smythers GW, Munroe DJ, Lemkin PF, TMAP poster, Swiss Proteomics Meeting, Geneva, Dec, 2002.

- Medjahed D, Smythers GW, Powell DA, Stephens RM, Lemkin PF, Munroe DJ, VIRTUAL2D: A Web-accessible predictive database for proteomics analysis, *Proteomics*, 2003, (in press, Feb).

- Medjahed D, Luke BT, Tontesh TS, Smythers GW, Munroe DJ, Lemkin PF, "TMAP" (Tissue Molecular Anatomy Project), an expression database for comparative cancer proteomics. *Proteomics*, 2003, (in press, June).

References