Mayday 2.9

Introductory Guide

Florian Battke

Contents

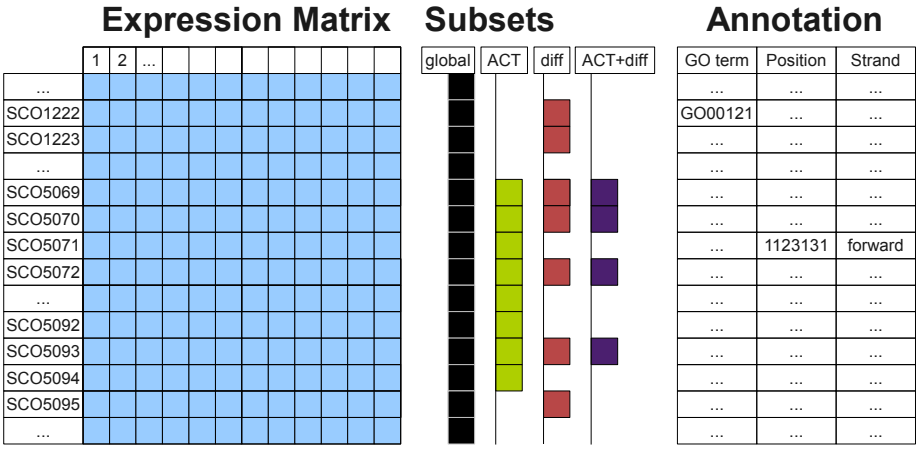


Figure 1: MAYDAY's main data structures, conceptual overview

1 Concepts

MAYDAY operates on numerical matrices. Each column of the matrix describes one sample under study (a tissue, a condition, a treatment, etc.) while each row describes an element under study (a gene, a protein, a genomic region, etc.). Cell (i, j) in the matrix contains the measured value for element i in sample j. Columns are called *experiments* and rows are called *probes* in MAYDAY. Most of MAYDAY's methods operate on probes.

During analyses, one of the most common tasks is to find and analyze groups of probes with similar characteristics. Groupings can be based on *a priori* information (e.g. genes belonging to a common pathway) or on derived values (e.g. probes with a certain level of variance in their measurements). MAYDAY groups probes into *ProbeLists*[1]. ProbeLists can be grouped together to form a hierarchy, where the "parent" list is the union of all lists contained in it.

Aside from the *primary* data in the matrix, additional information is usually available and of great use during analyses. These *meta information* values can be thought of as additional columns in the matrix, containing e.g. statistical significance values, genomic coordinates, pathway information, etc. Each additional column has a certain data type (numerical, text, genomic location, etc.) and may only contain values for a subset of the matrix rows, i.e. the columns are usually sparse. Figure 1 shows MAYDAY's core data structures.

MAYDAY organizes these additional columns as *meta information (MI) groups*, associating probes with *meta information objects (MIOs)*. MI Group can be organized hierarchically to improve ease of access for the user.

[1] In mathematical terms, ProbeLists are subsets of the matrix rows and need not be disjoint

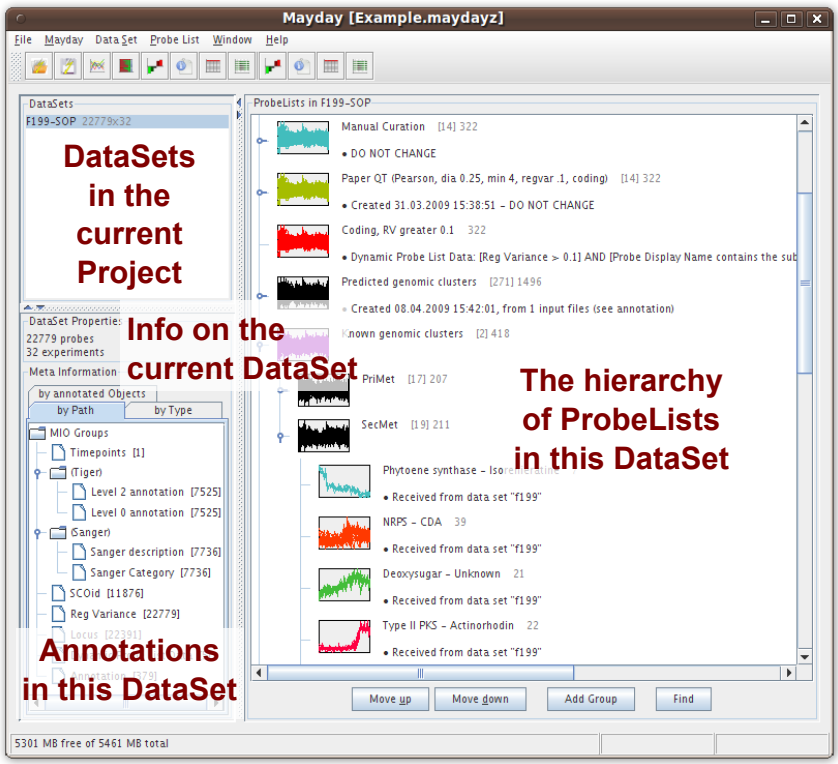


Figure 2: MAYDAY's main window with explanations

The primary and meta data together with the hierarchy of ProbeLists constitute a *DataSet*. Multiple DataSets can be opened at the same time in MAYDAY. Several datasets together form MAYDAY's largest unit of data, a *Project*, which can be stored in the efficient MAYDAY snapshot file format.

2 Mayday's main window

MAYDAY's main window is organized into three major parts (see figure 2). In the top left corner, the list of datasets in the current project is shown together with the dimensions of their primary data matrices. Actions can be performed on datasets by selecting them from the list and right-clicking, or selecting an action from the **DataSet** menu. The selected dataset[2] is used to fill the remainder of the main window:

In the lower-left corner, the hierarchy of meta information groups is displayed (the number of annotations in each group is indicated in brackets). Actions on meta-information groups can be performed by selecting groups and

[2]the last selected dataset if several are selected

right-clicking, or by opening the *Meta information manager* window from the **DataSet** menu.

The right part of the main window shows the hierarchy of ProbeLists in the current DataSet. The ProbeList name is followed by the number of subordinated lists (if any, in square brackets) and the number of probes contained in the list. Annotations are shown below the name, a preview plot of the list's contents is drawn as well (see section 4 for details). Actions on ProbeLists can be performed by selecting lists and right-clicking, or by choosing an action from the **ProbeList** menu.

3 Loading Data into Mayday

3.1 Projects

MAYDAY projects can be loaded and saved from the **File** menu. To add datasets to a project, open the project first and then import more datasets (see below).

3.2 Importing external data

To import data into MAYDAY, open the **DataSet** menu and select **Import from file**. Different file formats are supported. The most commonly used input format is the tabular format. When loading tabular data, a dialog window is shown where the column separator (e.g. tabulator, comma, etc) and other relevant parameters can be set. MAYDAY assumes that the first column of the matrix contains probe names. Probe names must be unique within each dataset. If the first row contains experiment names, the "Has header line" options must be selected. A content preview displayed in the dialog looks helps finding the right parameter combination (see figure 3).

A second dialog window is shown where each column can be assigned a data type. A column can either be ignored, imported as primary (experiment) data, as a ProbeList or as meta information. MAYDAY tries to automatically determine the type of each column. A ProbeList column may only contain one distinct value indicating, if present, the membership of the probe in the respective ProbeList. For meta information column, the desired type of the meta data (text, number, etc.) has to be selected (figure 3).

4 Managing ProbeLists

ProbeLists define subsets of rows of the matrix. Each ProbeList has a name that must be unique within the dataset. To find a specific ProbeList by name,

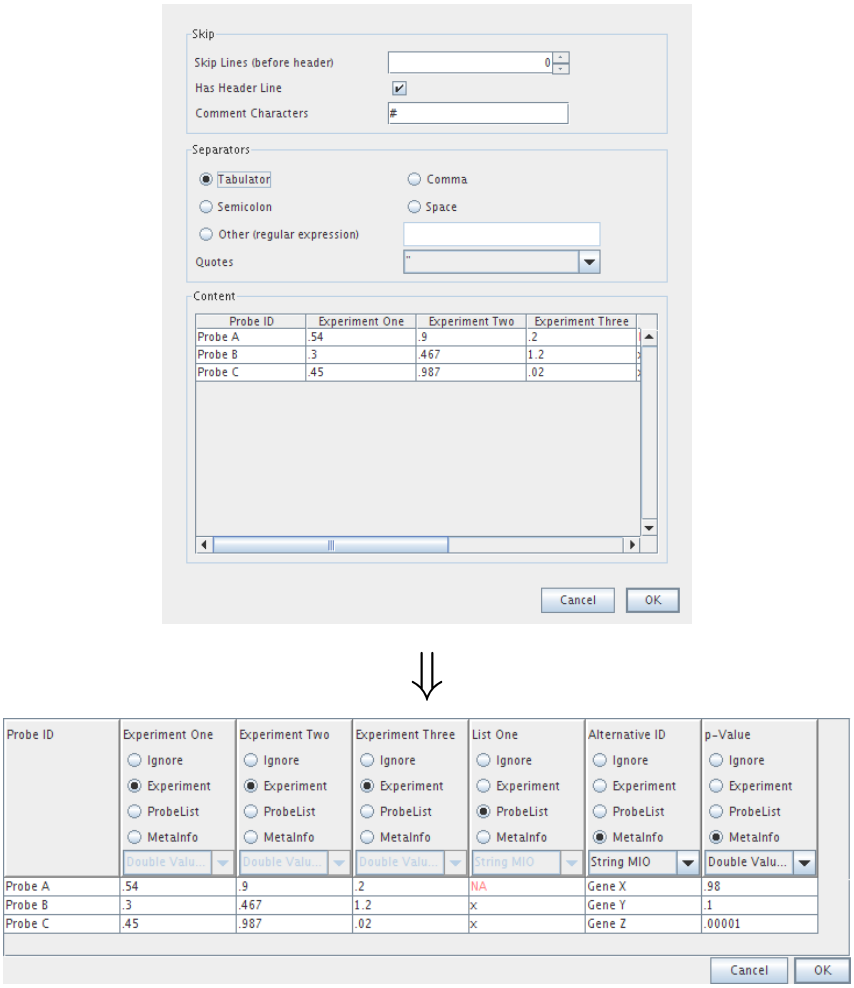


Figure 3: Importing tabular data into MAYDAY.

click the "Find" button or press CTRL-F. Furthermore, each ProbeList is associated a color that is used for the preview plot as well as for visualizations (see section 8). ProbeLists can be annotated with meta information, e.g. with text describing its provenance.

Double-clicking on the preview plot opens a larger visualizer, double clicking anywhere else on the ProbeList opens its properties window.

ProbeLists can be grouped together to form a hierarchy. Use the "Create Group" button in MAYDAY's main window to create a new group containing all currently selected ProbeLists. Use drag&drop to move ProbeLists to other groups or to change the order of ProbeLists within a group. Dragging a ProbeList onto a group of ProbeLists will insert it into the group. Dragged ProbeLists are always inserted *below* the ProbeList they are dropped on. The "Move up" and "Move down" buttons can be used to change the order of ProbeLists within a group.

ProbeLists can be exported to text files and imported from text files from the **ProbeList** menu. If several datasets are opened, ProbeLists can be sent from one dataset to another ("Further export option"→"Send to DataSet"). A new Dataset can be created from an existing ProbeList, the new DataSet will then only contain the probes found in that ProbeList.

If several ProbeLists are disjoint, they can be converted to categorical meta information (labelling each probe with the one probelist it was contained in), by selecting "to nominal MIO" from the "Further export options" submenu. To create ProbeLists from such nominal meta information, select "Create"→"from nominal MIO" from the **ProbeList** menu.

5 Meta Information

Meta information objects are additional values associated with probes. They are organized into Meta Information Groups, each group has a name, a fixed data type, and a position in the hierarchy of MI Groups. MAYDAY supports a large number of data types, among them numerical (integer and floating-point) values, textual (string) values, lists of strings and mappings of strings, boolean (true/false) values, genomic coordinates, etc.

5.1 Importing

As with the primary data, meta information can be imported from tabular files. To associate meta information with probes in the dataset, probe names and the identifiers given in the meta information file must be identical. Otherwise, a mapping can not be found. To import data from a tabular file, right-click in the meta information part of the main window, or open the Meta Information Manager from the **DataSet** menu. The steps necessary and the dialogs involved are the same as described in section 3.2.

Genetic coordinates can be imported from a number of file formats, among them tabular files, GFF and PTT as well as GenBank files, as long as there is a unique name associated with each coordinate. Choose "Add Locus information" from the context menu or within the Meta Information Manager.

5.2 Display names

In most microarray experiments, probe identifiers as provided by the array manufacturers' platforms are not very intuitive. Therefore, MAYDAY allows any textual meta information group to be used as *Probe Display Names*. If display names are set, they are used instead of the probe identifiers for visualization and display purposes. To use an existing meta information group as

display name group, select the group, right-click, and select "Use for probe display names". Display names can also be configured from the **DataSet** menu, submenu **Probe Names**.

If display names are not set, or if the display name group does not contain a value for a probe, the probe identifier is used instead. If display names are present, they can be used to map further meta information to the named probes.

5.3 Processing

Some plugins use meta information as input to produce new meta information. At present, there are a some such plugins that map numeric meta data to the interval $[0, 1]$ using different mapping functions.

5.4 Creating new meta information

Meta information is often created as a result of running MAYDAY plugins, e.g. to store p-values of a statistical test. These groups appear in the meta information window.

5.5 Using meta information in visualizations

Please see section 8 for a description on how to include meta information in visualizations.

6 Statistics

Statistical values are represented as meta information, either imported from external files as described above) or created by MAYDAY.

6.1 Simple statistical values

The **Statistics** submenu of the **ProbeList** menu contains a large number of methods to compute basic statistical variables for each probe, such as the mean or median expression. When selecting such a method, a new MI Group is added to the dataset.

6.2 Statistical tests

Statistical tests can be applied by choosing "Statistical Test" from the **Statistics** submenu of the **ProbeList** menu. This requires the name of a new

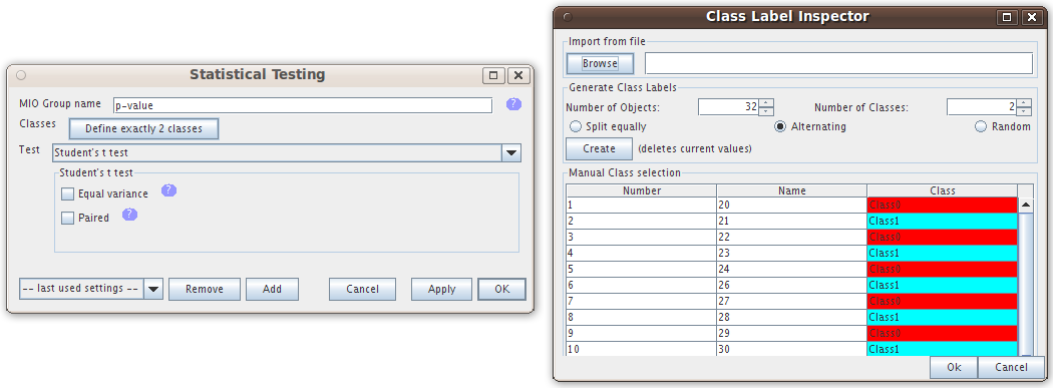


Figure 4: Statistical testing in MAYDAY.

MIGroup (to hold the resulting p-values), a class selection and a test statistic. Different statistics are available (t test, SAM, Rank Product, WAD), some of them offering further settings (see figure 4). After computing the test, one MI Groups will be created containing p-values. Some test statistics create additional MI Groups (hierarchically placed below the p-value group) containing the values of the test statistic (e.g. t scores).

7 Filtering

To create a new ProbeList using filtering rules, select "Filter..." from the **ProbeList** menu. This will create a new *dynamic* ProbeList. While regular ProbeLists contain a number of probes and allow probes to be added and removed at a later time (**ProbeList** menu, "Content"), dynamic ProbeList do not allow manual changes to their content. Instead, a set of rules is used to decide which probes to add to the list.

By default, the dynamic ProbeList contains rules to add all probes of all ProbeList that were selected when it was created. The Rule Editor (figure 5) can be used to change the rules of the dynamic ProbeList. Rules are built from processing modules that are linked together to form a processing chain. Each module accepts a certain type of input and returns another type as output. The input to the complete chain is a probe, the output a boolean value signifying whether the probe should be added to the list. Any number of rules can be combined with the AND and OR operators. The number of probes matching the criteria is updated automatically. By clicking "Apply" or "OK" in the dialog, all open plots showing the dynamic ProbeList are updated.

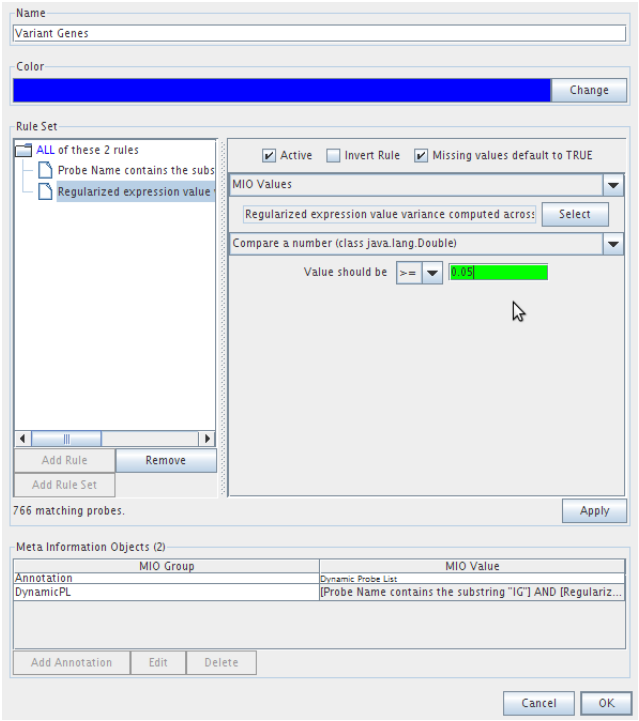


Figure 5: The rule set editor for dynamic ProbeLists

8 Visualization

8.1 The View Model

All visualizations in MAYDAY are based on a common view model. Several plots can share a common view model, effectively presenting different perspectives of the same underlying data. Several view models can coexist.

A view model contains a number of ProbeLists and their ordering as defined by the DataSet (see below). It also provides plots with *primary* data from the DataSet and can modify this data on-the-fly ("Online data manipulation"). Furthermore, the view model stores the list of currently selected probes and communicates changes to this selection between all its plots.

8.2 The ordering of ProbeLists

ProbeLists are not only associated with a position in a hierarchical structure but also with a unique index, which induces an ordering on the ProbeLists. The index corresponds to the position of the ProbeList in the right-hand side of MAYDAY's main window (assuming that the hierarchy is fully expanded and each ProbeList is visible): The top-most ProbeList has the smallest index, while the ProbeList at the bottom of the list has the highest index. Using

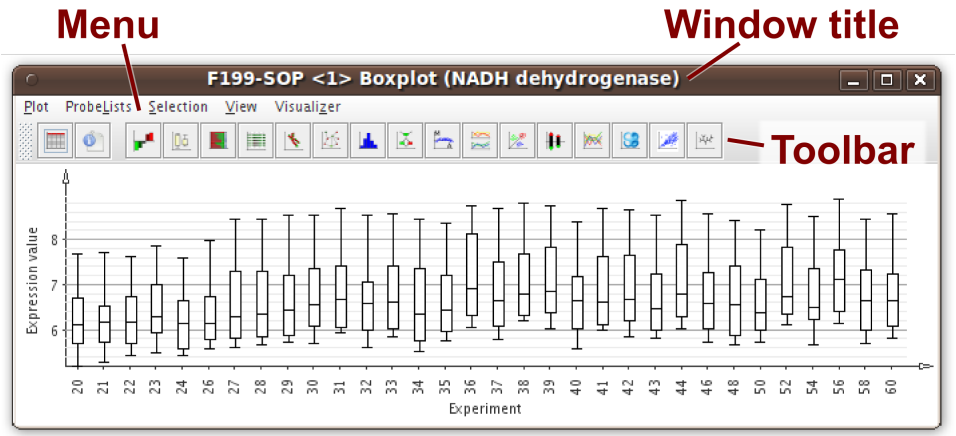


Figure 6: A visualizer window, see section 8.3 for a description.

the ordering induced by these indices, a *top-priority ProbeList* can be found for each probe in the view model, defined as the ProbeList with the smallest index among those ProbeLists containing said probe. This is important to understand because most plots offer coloring according to the top-priority ProbeList's color.

8.3 Visualizer window management

The windows for all plots sharing a view model are managed by a *visualizer*. Figure 6 shows an example plot window and will be used for the following discussion.

The title of each window is built from several parts, displaying

DataSet <Visualizer ID> Plot Title (ProbeLists)

As there can be several view models active for each DataSet, the window title shows the identifier of the current visualizer resp. view model.

Each window has a menu and a toolbar. Parts of the menu are the same for all plots, namely the **ProbeLists**, **Selection** and **Visualizer** menus. The first menu will either be **Plot** in the case of graphical plots or **Table** in the case of tabular views, and offers export methods. The **View** menu holds options relevant to the specific plot displayed in the window. The other menus will be presented now.

- **ProbeLists**
  Each view model contains a number of ProbeList. You can add or remove ProbeLists from the model, updating all plots accordingly.

- **Selection**
  Each view model manages a selection of probes. If probes are selected by interacting with any of the view model's plots, the selection is communicated to all other plots of that view model. The menu offers methods to change that selection:

  - Clear – removes all probes from the selection
  - Invert – sets all selected probes to unselected and vice-versa
  - Add probe – adds a single probe to the selection

  Furthermore, the current selection can be used to create or modify ProbeLists:

  - Create ProbeList – creates a new ProbeList from the current selection.
  - Create ProbeList Bipartition – creates one ProbeList for all selected probes and one for all unselected probes.
  - Remove from ProbeList(s) - Removes the currently selected probes from all ProbeLists in the view model. As a consequence, the probes will no longer be shown in any of the model's plots.
  - Send selection to ProbeList – Adds all selected probes to an already existing ProbeList

  There are also actions that interact with other visualizers:

  - Plot in separate visualizer – Creates a temporary ProbeList from the selection and plots it in a new view model.
  - Send selection to visualizer – Communicates this view model's selection to another view model.
  - Synchronize selection with visualizer – Links two visualizers together. As a result, any change to a selection in this (the *source*) visualizer is communicated directly to the other (*target*) visualizer. The link can be broken by selecting **Selection**→"Stop synchronizing selection" from the source visualizer.

  Finally, the "Probe" submenu offers further actions that work on the set of currently selected probes.

- **Visualizer**
  This menu contains further activities for the view model and its visualizer: The data manipulation can be changed, changing the primary values of all probes and updating plots accordingly. All plot windows associated with the visualizer can be brought to front or closed. Further

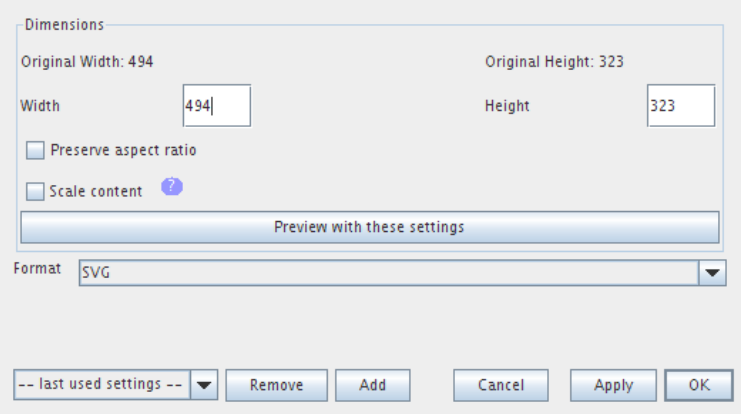


Figure 7: Settings dialog for graphics export

plots can be added to this visualizer, providing new perspectives on the same view model. The list of plot windows sharing this view model is also added to the menu. Finally, the "combine all" item merges all currently open plots of this visualizer into one window, useful for creating combined figures for publications.

Toolbar
The window toolbar contains buttons for all available plots. Clicking on any of these buttons opens a new plot window associated with the current view model, providing a new perspective on its data.

8.4 Exporting plots

All graphical plots can be exported as either bitmapped images (PNG, JPEG and TIFF formats) or as scalable vector graphics (SVG format) by either choosing "Export" or "Export visible area" from the **Plot** menu.

Figure 7 shows the export settings dialog. The dialog shows the current size of the plot in screen coordinates. Users can enter new dimensions for exporting and select whether they want the aspect ratio to remain the same (i.e. the ratio of width/height should not change) and how the plot should be adapted to the new dimensions.

The meaning of the "Scale content" setting is as follows (see figure 8 for an illustration):

- If "Scale content" is *disabled*, the plot will be drawn with the desired dimensions as if the plot window were resized to that dimension. Distances between plot elements are changed, the elements remain at the same size. As a result, axes might change to accommodate the new dimensions (i.e. adding or removing axis tick lines depending on available space). The

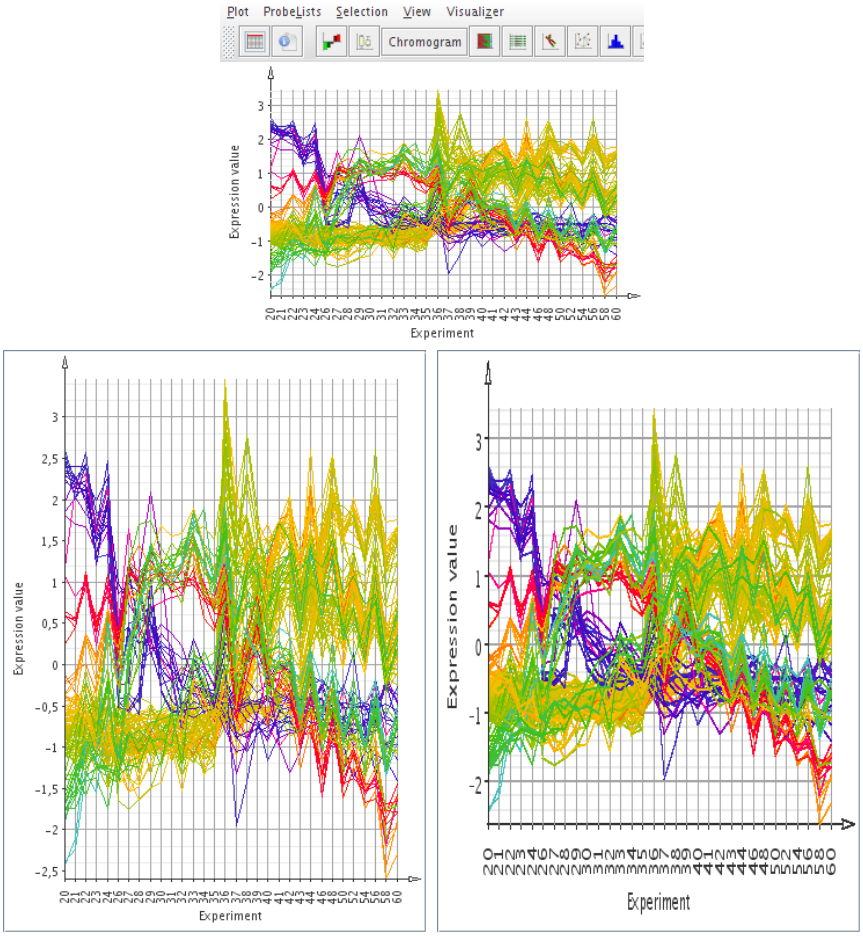


Figure 8: Illustration of the "Scale content" export setting. Disabling the setting changes plot dimensions (left) while enabling it stretches the plot to fill the desired output area (right). The original plot window is shown at the top, the plot height was doubled for export while the width was kept constant.

proportions of text will be unchanged.
This is the recommended setting for exporting to vector graphics files.

- If "Scale content" is *enabled*, the plot will be drawn with its original dimensions and then scaled to the desired output dimension by enlarging/shrinking all elements and distances between them by the same factor. If the aspect ratio of the output is different than that of the original plot, elements such as text will be appear stretched in one dimension. This setting (in combination with keeping the aspect ratio constant) is useful to create a high-resolution bitmap images of plots.

All bitmap file formats offer *anti-aliasing*. If enabled, lines and text will appear smoother (no visible "stairs" in diagonal lines).

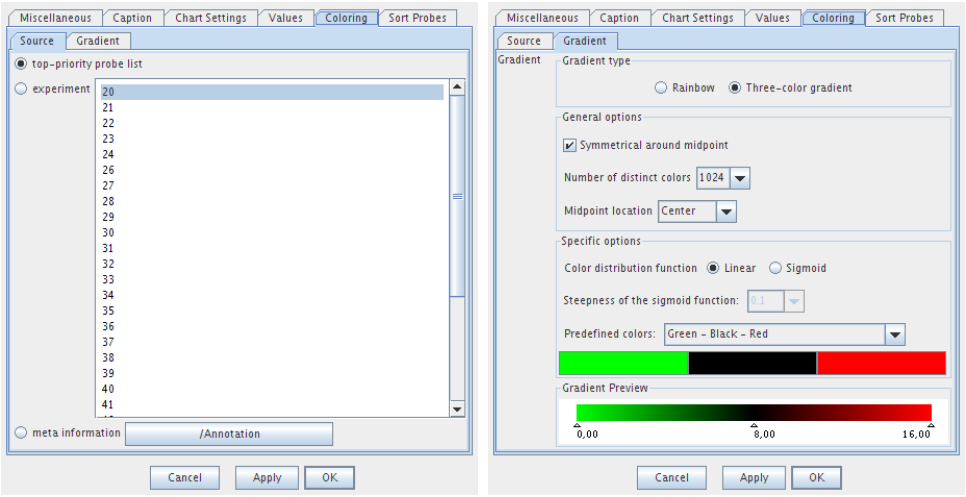


Figure 9: Mapping probes to colors using a color provider.

8.5 Coloring

MAYDAY uses a uniform framework for assigning colors to probes. There are three coloring methods (see figure 9 left):

1. **Color by top-priority probe list** – Each probe is assigned the color of its top-priority ProbeList as described in section 8.2.

2. **Color by experiment** – Each probe is represented by one value of its associated primary information, mapped to a gradient (see below).

3. **Color by meta information** – Each probe is represented by a value of its associated meta information. If the meta information is numeric, it is mapped to a gradient (see below). If the value is of a categorical type (a text string for instance), each distinct value is assigned a color automatically such that colors are as distinct as possible.

For numerical values, coloring is a two-step process. After each probe is translated into a numeric value (from primary or meta information as described above), the second step is mapping the numeric value to a color using a color gradient. There are a number of predefined gradients and users can create their own gradient by selecting three colors (see figure 9 right). The gradient can either be linear or sigmoid with a configurable steepness.

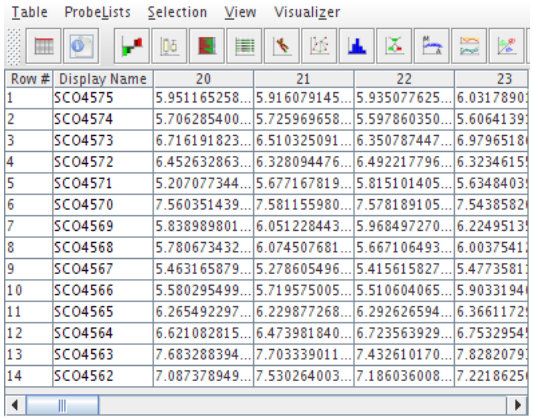
The choice of the "midpoint" value influences the mapping of values to colors. It specifies which value is mapped to the middle color. The choices are

- **Center** – The middle value ($\frac{\max - \min}{2}$) is mapped to the middle color.
- **Minimum** – The minimal value is mapped to the middle color.

- **Maximum** – The maximal value is mapped to the middle color.
- **Zero** – The value zero (0.0) is mapped to the middle color.
- **Custom** – The specified number is mapped to the middle color.

In the "custom" and "zero" cases when that value differs from the middle value, an asymmetrical gradient will use all available colors by compressing/stretching the interval left and right of the midpoint as needed. A symmetrical gradient will not compress/stretch intervals but will not be able to use all colors.

8.6 Mayday's tabular views



Row #	Display Name	20	21	22	23
1	SCO4575	5.951165258...	5.916079145...	5.935077625...	6.0317890
2	SCO4574	5.706285400...	5.725969658...	5.597860350...	5.6064139
3	SCO4573	6.716191823...	6.510325091...	6.350787447...	6.9796518
4	SCO4572	6.452632863...	6.328094476...	6.492217796...	6.3234615
5	SCO4571	5.207077344...	5.677167819...	5.815101405...	5.6348403
6	SCO4570	7.560351439...	7.581155980...	7.578189105...	7.5438582
7	SCO4569	5.838989801...	6.051228443...	5.968497270...	6.2249513
8	SCO4568	5.780673432...	6.074507681...	5.667106493...	6.0037541
9	SCO4567	5.463165879...	5.278605496...	5.415615827...	5.4773581
10	SCO4566	5.580295499...	5.719575005...	5.510604065...	5.9033194
11	SCO4565	6.265492297...	6.229877268...	6.292626594...	6.3661172
12	SCO4564	6.621082815...	6.473981840...	6.723563929...	6.7532954
13	SCO4563	7.683288394...	7.703339011...	7.432610170...	7.8282079
14	SCO4562	7.087378949...	7.530264003...	7.186036008...	7.2218625

MAYDAY has two tabular viewers, one displaying the expression matrix, i.e. the primary data associated with each probe, the other displaying the meta information for each probe. Both viewers can sort probes by primary or meta information, by name, display name or top-priority probelist. Probes can be searched by name or display name. Selection is possible by clicking, shift-clicking or control-clicking on any table cell. Probe identifiers can be colored as described in section 8.5. The tables can be exported.

8.7 Mayday's visualizations

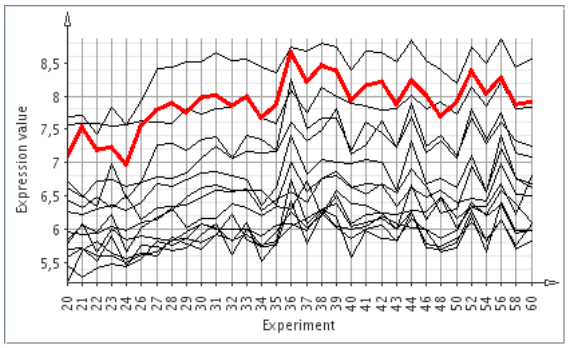
MAYDAY offers a large number of plots and new plots are added with each new release. We will not discuss them in great detail, instead we give a short overview of possible interactions and settings for each plot. Some characteristics apply to all plots (if not otherwise stated), such as the handling of titles and legends, of **selections**, **zooming** and **contextual information**:

- Select elements by clicking with the left mouse button.
  Hold CTRL while clicking to add to or remove from the selection.
  Some plots support "Range selection": Hold SHIFT while clicking to select a range of objects.

- Some plots support "brushing" selection, i.e. you can drag a frame around an area of interest. There are several selection modes:
  - Hold down CTRL to add probes to the previous selection (union)
  - Hold down ALT to intersect the new selection with the previous selection (intersect)
  - Hold down ALT-CTRL to remove the new selection from the previous selection (minus)
- Selected elements are highlighted in a user-definable color (default: red).
- Zoom in and out by holding CTRL and turning the mouse wheel.
  Zoom vertically by also holding SHIFT (CTRL-SHIFT-Mouse Wheel).
  Zoom horizontally by also holding ALT (CTRL-ALT-Mouse Wheel).
- Right-click to get information about selected probes.
- Coloring each element according to the probe it represents (see 8.5).

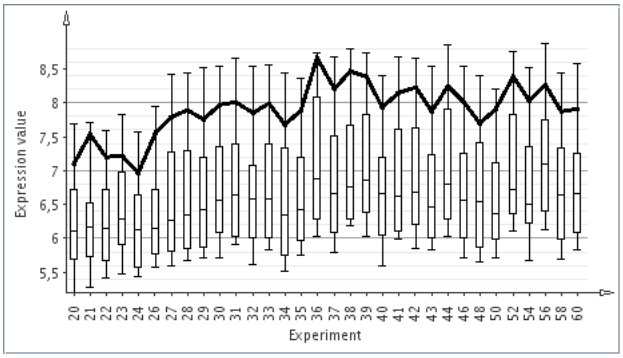
In the following sections we describe each visualization briefly.

8.7.1 Profile plot



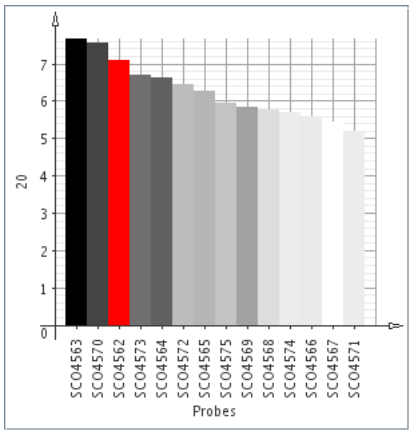
- Displays gene expression in parallel coordinates
- Profiles can be selected by clicking or by dragging a frame around an area of interest ("brushing").
- Supports time-points, i.e. the x axis can be scaled by time points instead of placing experiments at regular intervals.
- Introducing breaks: profiles can be disconnected at certain positions for enhanced clarity of exposition.

8.7.2 Box plot



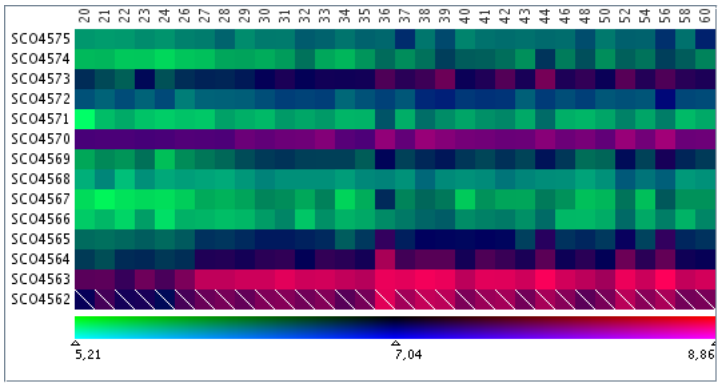
- Visualizes statistical properties (min, max, quartiles, median) for each experiment.
- Selection is not possible.
- Selected probes are displayed as profiles overlaying the boxes.
- Supports time-points, i.e. the x axis can be scaled by time points instead of placing experiments at regular intervals.

8.7.3 Bar plot



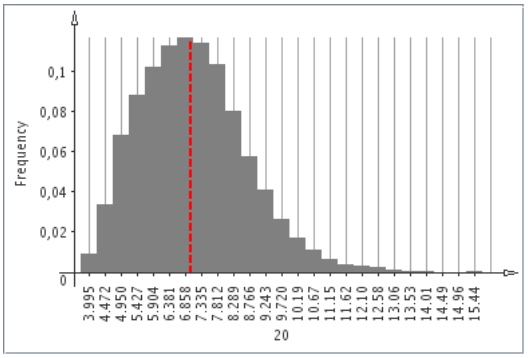
- Displays one bar for each probe.
- Selected probes are highlighted with the selection color.
- The value deciding the height of the bar can either be primary or meta information.
- Probes can be sorted by primary or meta information, by name, display name or top-priority probelist.

8.7.4 Enhanced HeatMap



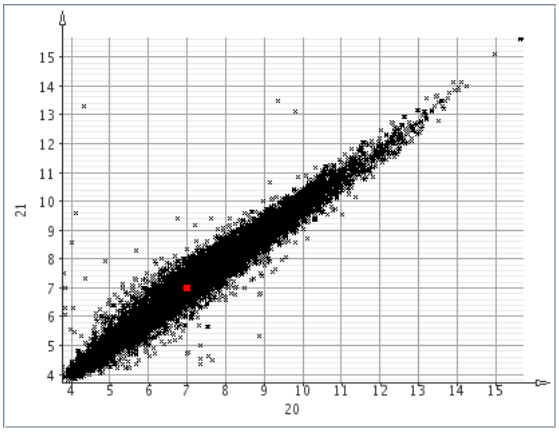
- Visualizes the expression matrix using colors.
- All selection methods are available, including range selection.
- Selected probes are displayed with hatched boxes.
- Probes (rows) can be sorted by primary or meta information, by name, display name, top-priority probelist or using a hierarchical clustering tree.
- Experiments (columns) can be sorted by name, index, or using a hierarchical clustering tree.
- Hierarchical clustering results can be added to the left (probe clustering) or above the plot (experiment clustering).
- The color gradient of the heatmap is independent of the probe identifier coloring.
- Enhancements using meta information:
  - overlaying a blue color
  - adding transparency
  - scaling the rows' heights
  - adding additional columns to the heatmap

8.7.5 Histogram



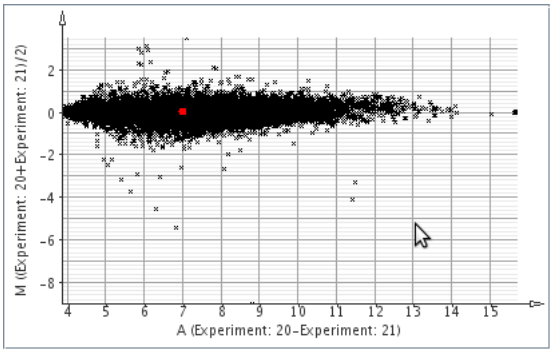
- Visualizes a distribution of values.
- Range selection is possible.
- Selected probes are indicated by a dashed line indicating their place in the distribution.
- The values plotted can either be primary data or meta information.

8.7.6 Scatter plot



- Visualizes the relationships between two values.
- Points can be selected by clicking or by dragging a selection rectangle around an area of interest ("brushing").
- Selected probes are highlighted with the selection color.
- The values plotted on each axis can either be primary or meta information.

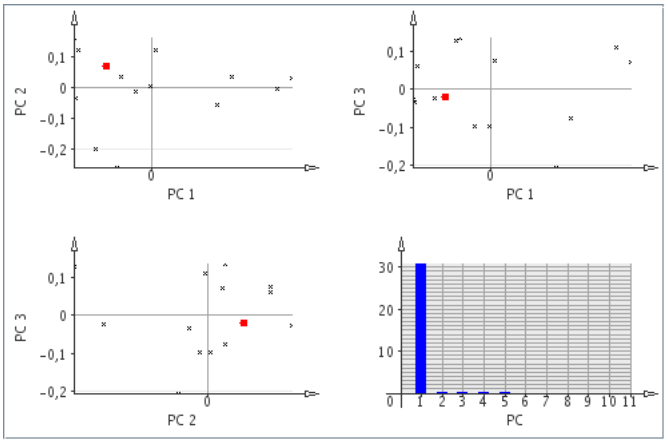
8.7.7 MA plot



- Visualizes the relationships between intensity (average intensity, $A = \frac{R+G}{2}$) and fold-change (difference in expression, $M = R - G$).

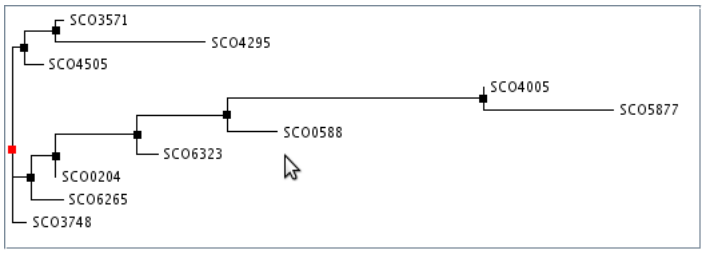
- Probes can be selected in any of the three PC plots either by clicking or by dragging a selection rectangle around an area of interest ("brushing").
- Selected probes are highlighted with the selection color.
- The data used to compute the M and A values can either be primary or meta information.

8.7.8 Principal component plot



- Computes a principal component analysis and displays the first three components against each other as well as the variance explained by each PC.
- Probes can be selected in any of the three PC plots either by clicking or by dragging a selection rectangle around an area of interest ("brushing").
- The PC matrix can be exported as a text file.

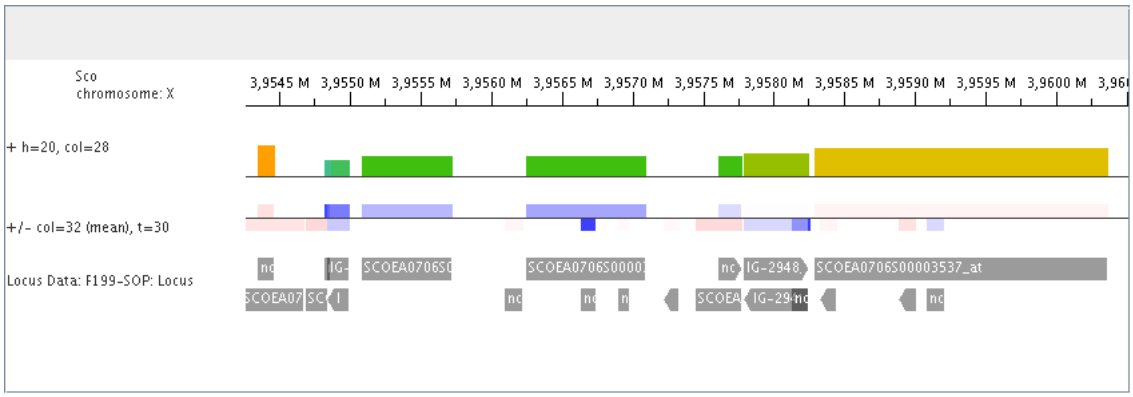
8.7.9 Tree visualizer



- Displays the tree resulting from a hierarchical clustering.
- Range selection is not supported.
- Selected probes are highlighted with light blue.
- The tree can be displayed using dendrogram, circular or unrooted layouts.

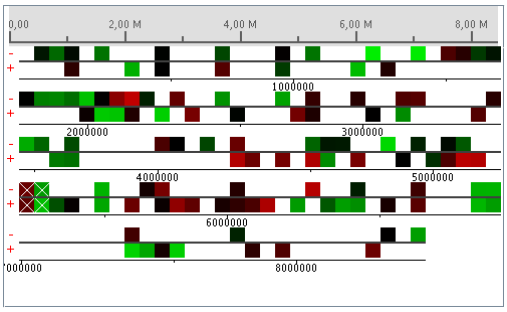
- The PC matrix can be exported as a text file.
- Right-clicking on a node allows tree editing operations (rooting, swapping children)
- The tree can be exported as newick file.
- If an experiment clustering is displayed, a selected edge can be used as a starting point for the gene mining plugin.

8.7.10 Genome Browser



- Track-based display of primary or meta information in a genomic context
- Selected probes are highlighted with blue triangles above/below tracks.
- A description of all possible interactions is given when the plot is first opened.

8.7.11 Genome Heat Stream



- Displays primary or meta information with genomic context as boxes in stacked rows.
- Selected probes are highlighted by hashed boxes.

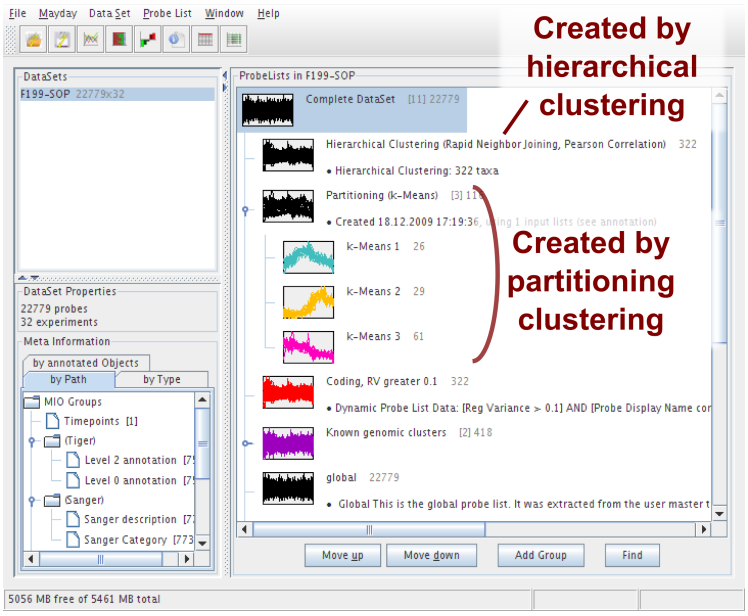


Figure 10: Partitioning clustering results in several new ProbeLists, hierarchical clustering creates exactly one new ProbeList.

9 Data Mining

The aim of data mining is to find patterns in data, e.g. gene expression profiles with common characteristics. There are several methods to *cluster* genes together based on their expression profiles, some of them are *hierarchical* (resulting in clusters that embedded in other clusters), others are *partitioning* (producing non-overlapping clusters). MAYDAY supports several methods for both types of clustering which can be started from the **ProbeList** menu, submenu **Data Mining**, **Clustering**.

9.1 Partitioning Clustering

Partitioning clustering methods create new ProbeLists, each one corresponding to one of the clusters found (see figure 10). Four methods for are currently implemented in MAYDAY:

- **k-means clustering**
  The well-known k-means method requires as input the desired number of clusters (k) and a distance measure. Other parameters can also be set in MAYDAY but can mostly be left at the default values. If the number k is unknown, the "Find optimal k" plugin can be used. It computes clusterings for different values of k and displays a plot of the within cluster variance for each k. Users should try to find a 'kink' in

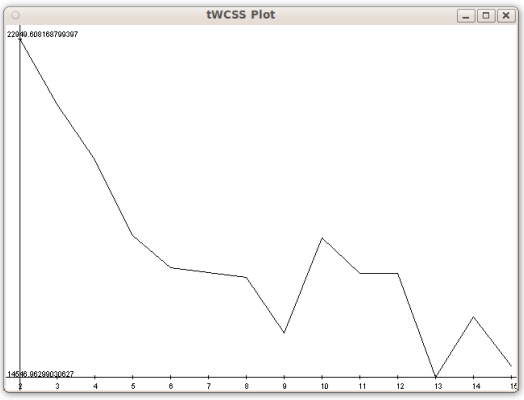


Figure 11: Finding the right k for k-means. The plot indicates that a good value would be $k = 6$.

the resulting plot, indicating that the clusters do not get much better even when using higher values of k (see figure 11).

- **QT-clustering**
  The QT clustering method [2] aims at finding clusters with a predefined quality. As input it requires the minimal cluster size, the diameter threshold and a distance measure. MAYDAY provides a method for finding the correct diameter threshold (see figure 12): The plot shows the distribution of distances between the probes. If there are clear clusters in the data, the distribution (red curve, top half of the window) should be bimodal. A good threshold is one lying between the two peaks. If the distribution appears unimodal, choosing a threshold around the middle of the rising (left) slope generally works well.

- **Density-based clustering**
  The DBScan algorithm [1] searches for clusters that are defined by a maximal distance between the points (probes). Probes that are too far away from other probes are clustered into a special "noise" cluster. It works particularly well on data with non-globular clusters. It requires as input the minimal number of points in each cluster and the maximum distance between two points in a cluster. Selecting the minimal number of points, a distance measure and a prefix for the resulting ProbeList names, the distribution of distance in the data is computed and diplayed. The plot can serve as a guide to finding the correct value for the maximal distance.

- **Self-organizing maps**
  MAYDAY provides a method to compute a two-dimensional SOM clustering [3]. Clusters in a self-organizing map are arranged in rows and

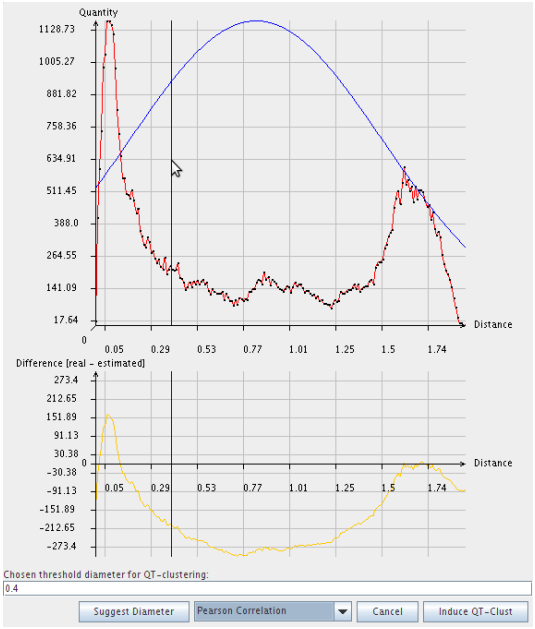


Figure 12: Finding the right diameter threshold for QT clustering, see the text for an interpretation.

columns and have a neighbor-relationship: The cluster in row 1, column 1 is more similar to the cluster in row 1, column 2 and the cluster in row 2, column 1, than to a cluster in row 7, column 6, for instance. A good way to visualize the outcome of a SOM clustering is to use MAYDAY's *multi-profile plot*, as shown in figure 13.

9.2 Hierarchical Clustering

When using an hierarchical clustering method, only one new ProbeList is created. It contains all probes used for clustering as well as the resulting clustering *tree* (see figure 10). This ProbeList can be used to visualize the tree alone (Tree Visualizer) or to display a heat map with the tree attached. MAYDAY offers different hierarchical clustering methods, the most widely used ones are UPGMA and Rapid Neighbor Joining [4]. With UPGMA, all leaves have the same distance from the root which is not the case for Rapid NJ.

These methods can be used to either cluster probes or to cluster experiment (transposed matrix).

Hierarchical clustering is a time- and memory-intensive task so input to these method should be filtered to reduce complexity (e.g. only clustering variant probes).

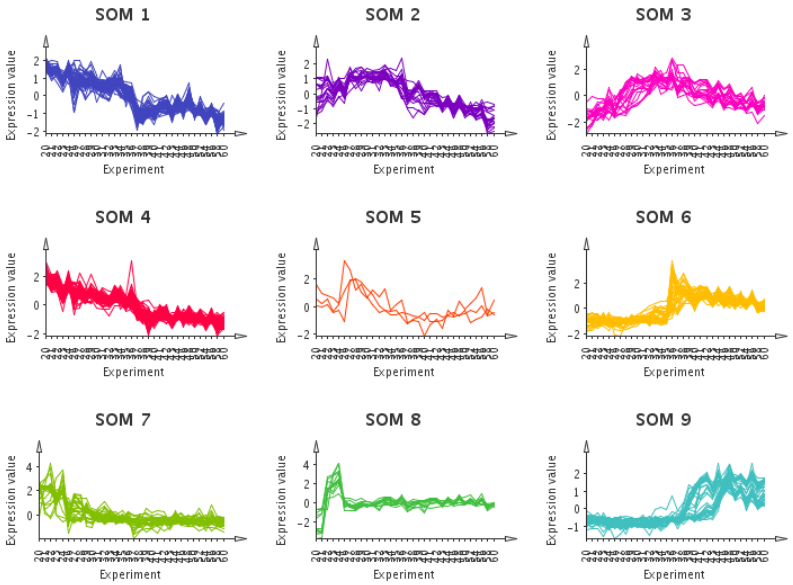


Figure 13: Self-organizing maps creates clusters with two-dimensional coordinates (row, column) that indiciate the closeness of the profiles, see section 9.1 for an explanation.

9.3 Gene Mining

The Gene Mining plugin can be used to find genes that are significantly differentially expressed between two classes of experiments with more sophisticated methods than e.g. a simple t test. We are currently working on the plugin's user interface to make it more intuitive and will add a description here when this work is done.

10 Configuration

MAYDAY's settings can be changed by selecting "Preferences" from the **Mayday** menu (figure 14). The "Appearance" tab allows to change the way previews are displayed (see section 8 for an explanation of data manipulations).

The plugin directory needs to be set correctly in the "Plug-Ins" tab so that MAYDAY can find all installed plugins. Make sure that the directory exists and is correct. The plugin directory must also be writable (i.e. the user must have permission to create and modify files in this directory) as some plugins use it to store additional files. To find out which plugins have been found, select "Plugins" from the **Mayday** menu.

Apart from MAYDAY's core preferences, some plugins also offer configurable parameters. One of them is the "Toolbar" plugin, which allows users to add

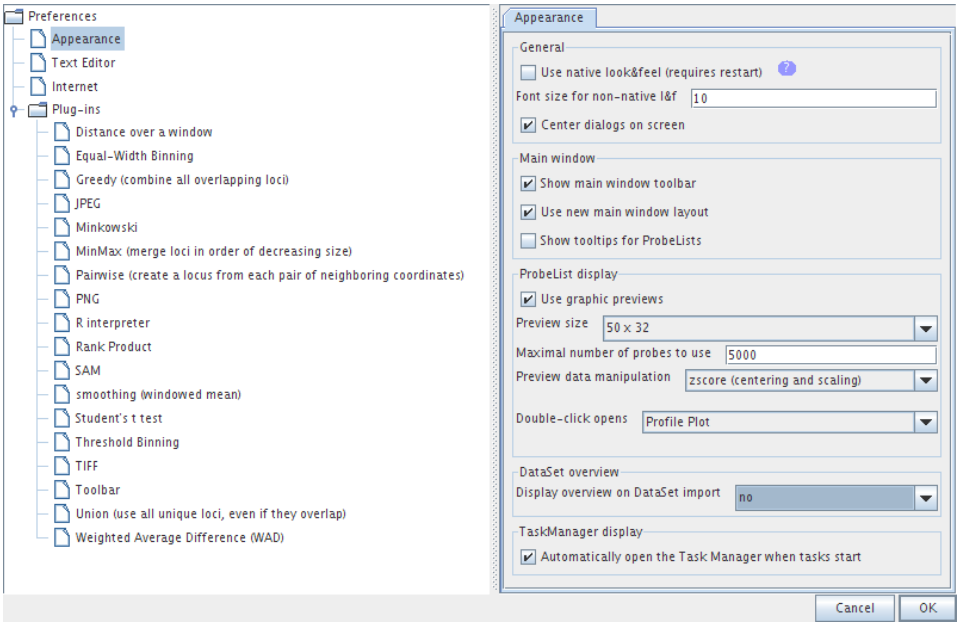


Figure 14: MAYDAY's preference dialog

often-used plugins to the main window toolbar (provided the toolbar is activated in the "Appearance" tab.

11 Running Mayday with more memory

Casual users will most likely run MAYDAY via the WebStart facility. However, memory available to WebStart applications is limited and can not be changed easily by the user. If your analyses require more than the default 512 MB of memory, consider installing MAYDAY locally and changing the amount of memory assigned to the Java virtual machine. Bear in mind that Java can only handle 2GB of memory when running on a 32 bit operating system.

If you have a 64 bit operating system, please make sure that you are running the correct Java version. Download and extract the MAYDAY package from our website. The package contains a script file (`mayday.sh` for Linux/MacOS, `mayday.bat` for Windows) which can be used to start MAYDAY. To change the amount of memory available for MAYDAY, edit this file and change `-Xmx1G` to reflect your whishes (e.g. `-Xmx8G` for 8 gigabytes of memory). You should not assign more memory to MAYDAY than is physically available on your machine because computations performed on swapped memory are extremely slow.

12 Tracking bugs and getting help

If MAYDAY does not work as expected, a message window can be opened from the **Help** menu. Additional information may be presented there. If you think you have found a bug, please send us an email including a description of the problem, your system and the contents of the message window, or file a bug report at our bug tracker (also available from the **Help** menu).

13 Further documentation

Several plugins have their own documentation which can be found on the MAYDAY website:

Plugin Name	Description
RLink	An interface to the R statistical environment with a user-friendly terminal
MachineLearning	Building and applying classifiers, feature selection
MPF	Building pipelines and applying processing pipelines
Pathways	Pathway visualization

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[2] L J Heyer, S Kruglyak, and S Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*, 9(11):1106–15, Nov 1999.

[3] Teuvo Kohonen. *Self-Organizing Maps*. Springer New York, 1997.

[4] Martin Simonsen, Christian N.S. Pedersen, and Thomas Mailund. Rapid Neighbor-Joining. In *Proceedings of the 8th Workshop on Algorithms in Bioinformatics (WABI 2008)*, 2008.