BioNumerics Tutorial:

Clustering a binary data set

1 Aim

Cluster analysis is a collective noun for a variety of algorithms that have the common feature of visualizing the hierarchical relatedness between samples by grouping them in a dendrogram or tree.

In this tutorial we will create a dendrogram based on a binary data set, i.e. a data set with only two possible output values. We will specify the settings related to the similarity coefficient for calculation of the similarity matrix and the clustering method to be applied. We will also see how to alter the layout of the dendrogram and how to export the cluster analysis to use it in a publication, presentation, etc.

2 Preparing the database

1. Import the binary dataset from the example Excel file `Binary_data.xlsx` in the **DemoBase Connected** as described in the tutorial: "Importing non-numerical character data" or "Importing binary character data".

2. Click on a green colored dot in the *Experiment presence* panel to open the **Genes** experiment card for an entry (see Figure 1).

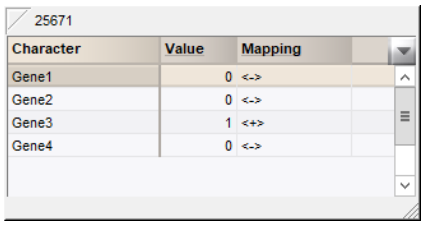

Figure 1: The character experiment card.

The character values (0 and 1) are displayed in the **Value** column.

3. Close the experiment card by clicking in the left upper corner of the card.

3 Comparison window

1. In the *Database entries* panel of the *Main* window, select all entries that have an associated **Genes** experiment: right-click on the header of the **Genes** column in the *Experiment presence* panel (i.e. the middle panel of the *Main* window) and select ***Select entries with experiment***. Alternatively select all entries with **CTRL+A** and unselect the entries defined as STANDARD.

2. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** () to create a new comparison for the selected entries.

3. Click on the  next to the experiment name **Genes** in the *Experiments* panel to display the **Genes** data in the *Experiment data* panel (see Figure 2).

Initially, the character values are displayed as colors according to the color scale defined for each character.

4. Select ***Characters*** > ***Show values+colors*** ( ) to display the corresponding character values and colors in overlay.

5. Select ***Characters*** > ***Show mappings+colors*** ( ) to show the mapping values and colors in overlay (if defined in the character type experiment).

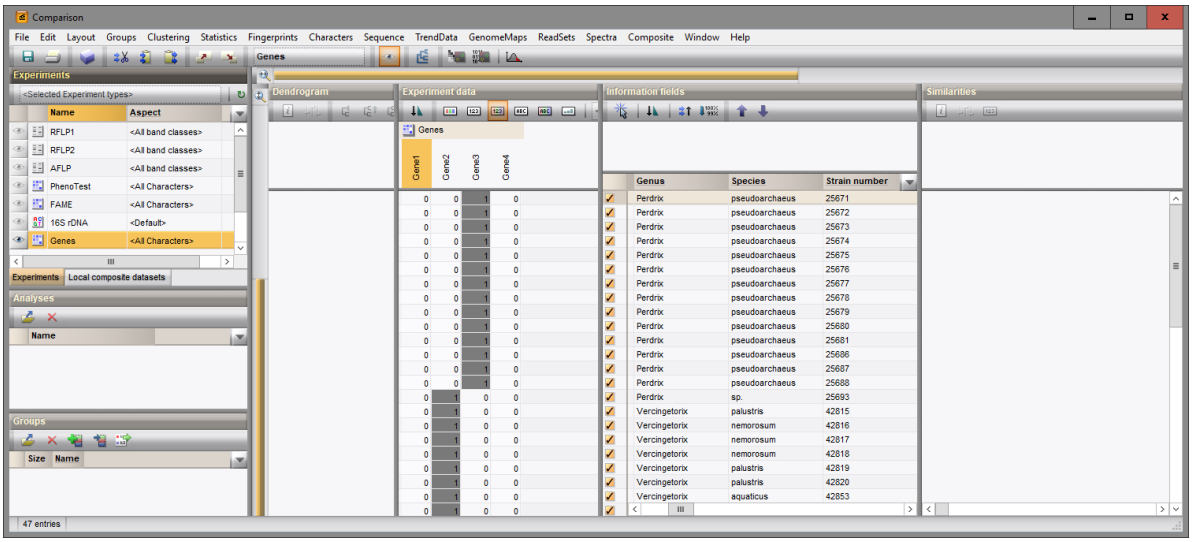

Figure 2: The *Comparison* window.

4 Cluster analysis

Cluster analysis is a two-step process. First, all pairwise similarity values are calculated with a **similarity coefficient**. Then, the resulting similarity matrix is converted into a dendrogram with a **clustering algorithm**. Although in practice these steps are performed together, they each require their own comparison settings.

1. Make sure **Genes** is selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

The first step deals with the similarity coefficient for the calculation of the similarity matrix. Depending on the selected coefficient, the relevant settings are displayed on the right. The coefficients are subdivided in three categories: ***Numerical***, ***Categorical***, and ***Binary***.

2. Since our data set only contains two states, select a ***Binary*** coefficient from the list, e.g. ***Jaccard correlation*** and press <***Next***>.

In step two the options related to the clustering algorithms are grouped. Under ***Method***, the clustering algorithm to be applied on the similarity matrix can be selected. A ***Dendrogram name*** can be entered in the corresponding text box. By default, the name of the experiment type appended with the aspect (here: "Genes(<All characters>)") will be used.

3. Make sure ***UPGMA*** is selected and press <***Finish***> to start the cluster analysis.

During the calculations, the program shows the progress in the *Comparison* window's caption (as a percentage), and there is a green progress bar in the bottom of the window.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The

cluster analysis is listed in the *Analyses* panel of the *Comparison* window.

4. Press the **F4** key to clear any selection in the database.

A branch can be moved up or down to improve the layout of a dendrogram:

5. Click the branch which you want to move up in the dendrogram and select ***Clustering*** > ***Move branch up*** ( ).

6. Click the branch which you want to move down in the dendrogram and select ***Clustering*** > ***Move branch down*** ( ).

Comparison groups can be defined from clusters, from database fields, or just from any selection you want. As an example, we will let BioNumerics create groups based on the **Genus** names.

7. In the *Comparison* window, right-click on the field name **Genus** in the *Information fields* panel, and select ***Create groups from database field***.

8. Keep the first option selected and confirm.

In our example three groups are created. The groups are listed in the *Groups* panel. The group color is displayed next to each entry in the *Information fields* panel (see Figure 3).

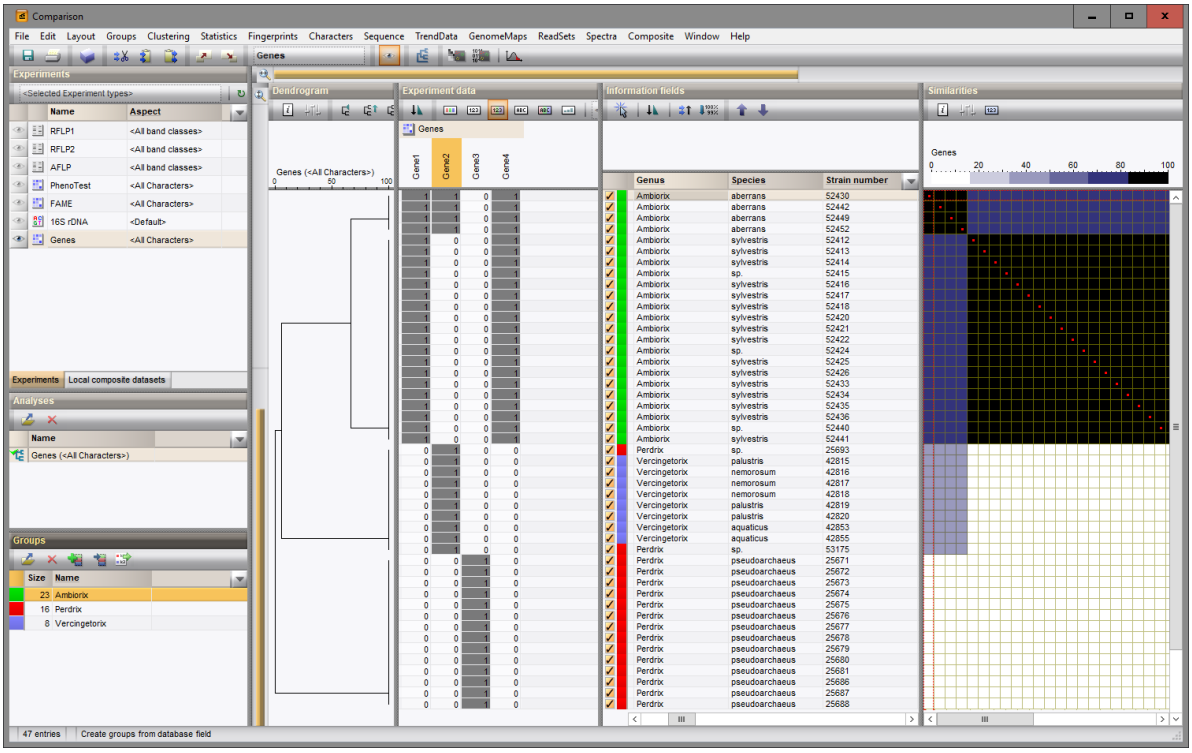

Figure 3: The *Comparison* window with groups defined.

The similarity values in the *Similarities* panel are represented by shades of blue.

9. To show the values in the matrix, select ***Clustering*** > ***Similarity matrix*** > ***Show values*** ( ).

10. Save the comparison with the dendrogram by selecting ***File*** > ***Save*** (, **Ctrl+S**). Specify a name and press <***OK***>.

11. Close the comparison with ***File*** > ***Exit***.

5 Exporting and printing a cluster analysis

BioNumerics can export the cluster analysis as it appears in the *Comparison* window.

1. Select ***File*** > ***Print preview...*** (, **Ctrl+P**).

The *Comparison print preview* window now appears.

2. To scan through the pages that will be printed out, use ***Edit*** > ***Previous page*** (, **Page Up**) and ***Edit*** > ***Next page*** (, **Page Down**).

3. To zoom in or out, use ***Edit*** > ***Zoom in*** (, **Ctrl+Page Up**) and ***Edit*** > ***Zoom out*** (, **Ctrl+Page Down**) or use the zoom slider.

4. To enlarge or reduce the whole image, use ***Layout*** > ***Enlarge image size*** () or ***Layout*** > ***Reduce image size*** ().

5. If a similarity matrix is available, it can be included with ***Layout*** > ***Show similarity matrix*** ().

6. On top of the page, there are a number of small yellow slider bars, which can be moved.

7. To preview and print the image in full color select ***Layout*** > ***Use colors*** ().

8. Export the image to the clipboard with ***File*** > ***Copy page to clipboard*** () and selecting an appropriate format.

9. If a printer is available, use ***File*** > ***Print this page*** () or ***File*** > ***Print all pages*** () to print one or all pages.

10. Select ***File*** > ***Exit*** to close the *Comparison print preview* window.