BioNumerics Tutorial:

Clustering a FAME 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 FAME data set. 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

The **DemoBase Connected** will be used in this tutorial and can be downloaded directly from the *BioNumerics Startup* window or restored from the back-up file available on our website:

1. To download the database directly from the *BioNumerics Startup* window, click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window. Select **DemoBase Connected** from the list and select ***Database*** > ***Download***. Confirm the download action.

2. To restore the database from the back-up file, first download the file `DemoBase_Connected.bnbk` from `http://www.applied-maths.com/download/sample-data`, under 'DemoBase Connected'.

   In the *BioNumerics Startup* window, press the  button, select ***Restore database***, browse for the downloaded file and select ***Create copy***. Specify a name and click <***OK***>.

In contrast to other browsers, some versions of Internet Explorer rename the `DemoBase_Connected.bnbk` database backup file into `DemoBase_Connected.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

3 Example data

The **FAME** data set used in this tutorial contains Fatty Acid Methyl Ester (FAME) profiles obtained on a Hewlett Packard 5890A gas-liquid chromatography instrument. This is a typical example of an *open* data set: the number of fatty acids found depends on the group of entries analyzed. If more entries are added, more fatty acids will probably be found.

1. In the *BioNumerics Startup* window, double-click on the **DemoBase Connected** database to open it.

2. Double-click on the **FAME** experiment in the *Experiment types* panel to open the *Character type* window (see Figure 1).

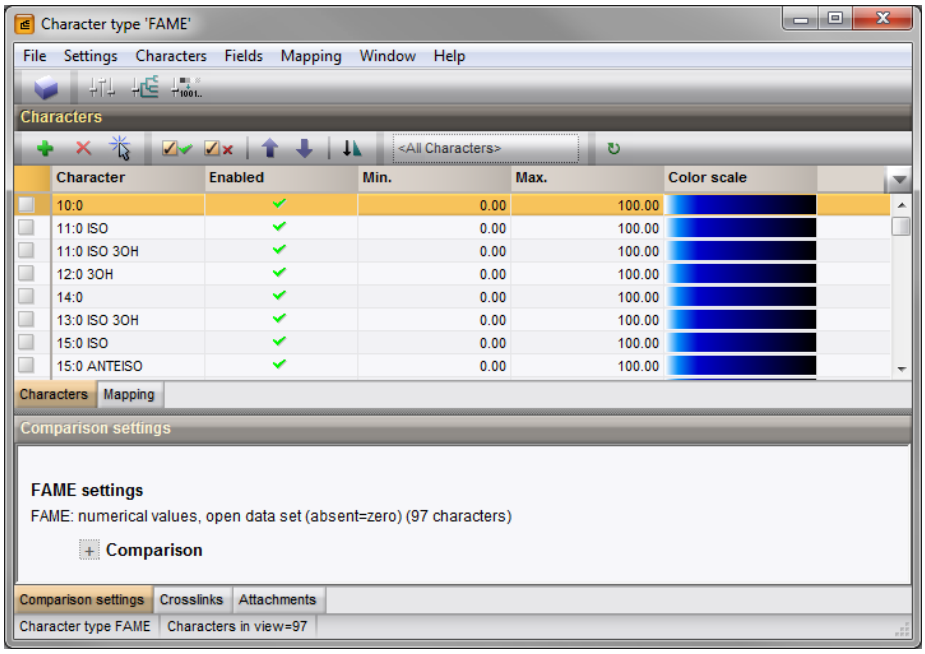

Figure 1: The *Character type* window.

The same *color scale* has been defined for all characters: white (lowest values) - blue (intermediate values) - black (highest values). The color scale makes it possible to assess character values at a glance in the *Comparison* window (see further).

The color scale is defined based on a *character range*. In the FAME data set the minimum value for all characters is set to 0 and the maximum value to 100.

3. Close the *Character type* window.

If a **FAME** experiment is present in the database for an entry, a green dot appears in the **FAME** column of the *Experiment presence* panel (i.e. the middle panel of the *Main* window).

4. Click on a green colored dot next to an entry, corresponding to the **FAME** experiment.

The *Experiment card* window lists all FAME profiles in the **Character** column and the values in the **Value** column (see Figure 2 for an example).

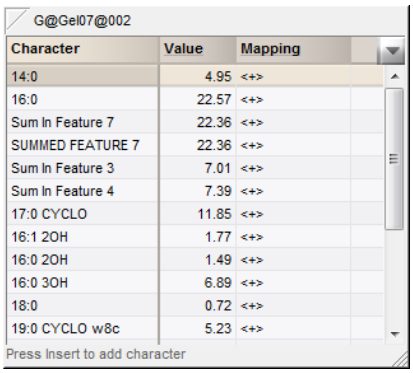

Figure 2: The *Experiment card* window of the FAME character experiment.

5. Close the *Experiment card* window by pressing the close button in the upper left corner of the card.

4 Comparison window

1. In the *Database entries* panel of the *Main* window, select all entries that have an associated **FAME** experiment: right-click on the header of the **FAME** 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 use the **CTRL-** key to 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 **FAME** in the *Experiments* panel to display the **FAME** data in the *Experiment data* panel (see Figure 3).

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

4. Select ***Characters*** > ***Show bar graphs*** ( ) to display the character values as colored bar graphs.

The colors used for the bar graphs are those as defined in the color scale for each character, while the bar heights are proportional to the character values, expressed as a percentage of the character range.

5. Select ***Characters*** > ***Show values*** ( ) to show the corresponding character values for all entries in the comparison.

6. The colors can also be shown in overlay with the values with ***Characters*** > ***Show values+colors*** ( ) (see Figure 3).

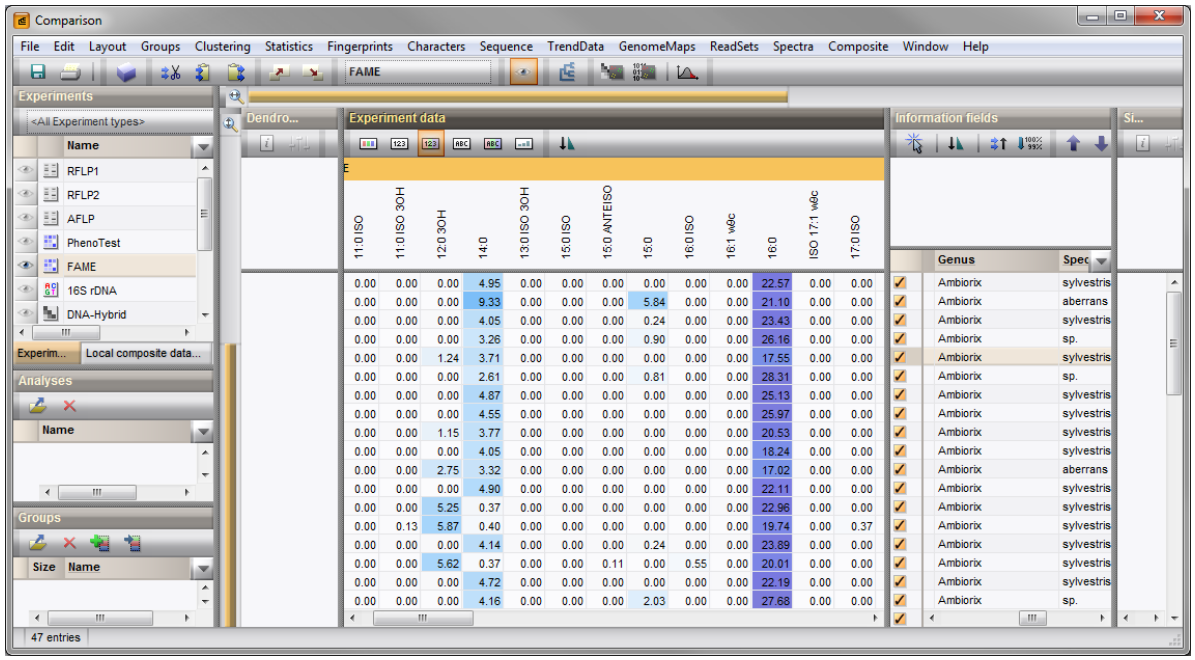

Figure 3: The *Comparison* window.

A convenient option to quickly check the behavior of an individual character is to list the entries according to the values of this character:

7. Select a **FAME** character in the *Experiment data* panel and select ***Characters*** > ***Sort by character value*** ( ).

The entries are sorted by increasing value of the selected character.

5 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 **FAME** 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.

2. Select ***Euclidean distance*** from the list 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: "FAME(<All characters>)") will be used.

3. Select ***UPGMA*** and <***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.

5. Left-click on the dendrogram to place the cursor on any node or tip (where a branch ends in an individual entry).

6. To select entries in a cluster, click on the node of the cluster while holding the **Ctrl**-key.

7. Press ***Edit*** > ***Cut selection*** ( , **Ctrl+X**) to remove the selected entries from the cluster analysis. Confirm the action. The dendrogram is automatically updated.

8. Select ***Edit*** > ***Paste selection*** ( , **Ctrl+V**). The cluster analysis is recalculated automatically, and the selected entries are placed back in the dendrogram.

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

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

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

To simplify the representation of large and complex dendrograms, it is possible to simplify branches by abridging them as a triangle.

11. Select a cluster of closely related entries and select ***Clustering*** > ***Collapse/expand branch*** (). Repeat this action to undo the abridge operation.

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.

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

13. 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 4).

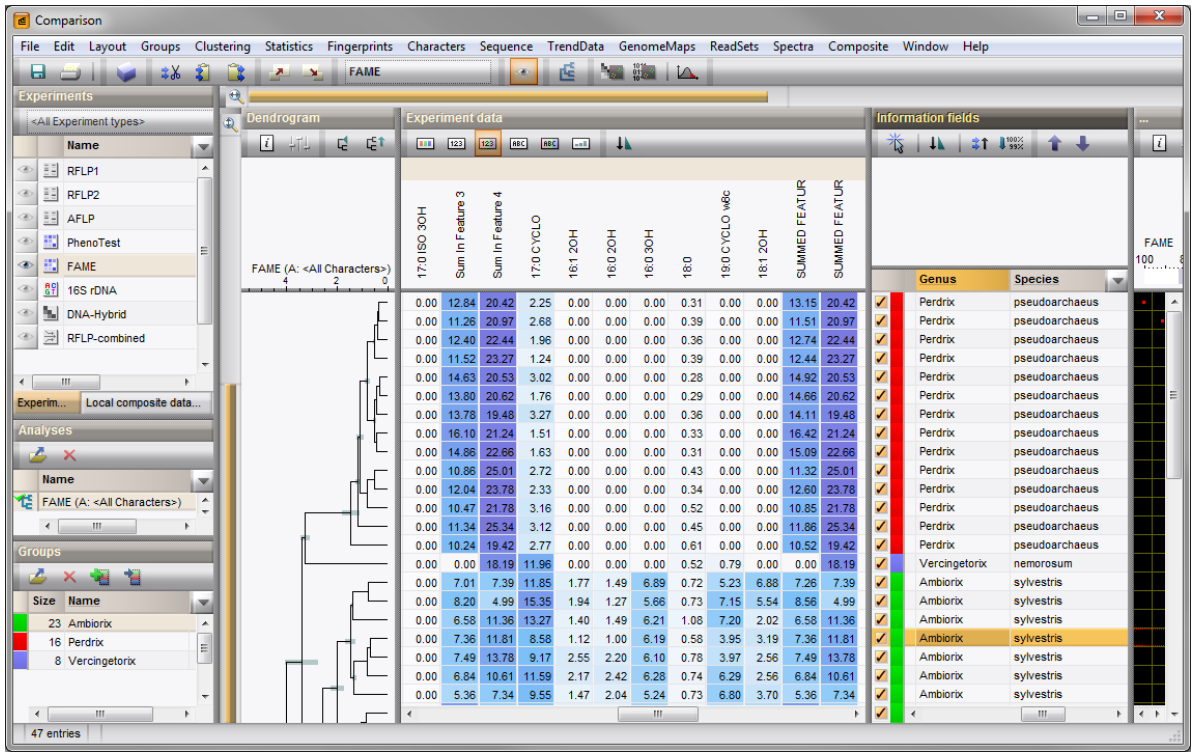

Figure 4: The *Comparison* window with groups defined.

14. Select ***Clustering*** > ***Dendrogram display settings...*** ( ) to call the *Dendrogram display settings* dialog box.

15. Enable ***Show group colors*** and press <***OK***>.

The dendrogram branches are now colored according to the group colors (see Figure 5).

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

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

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

6 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.

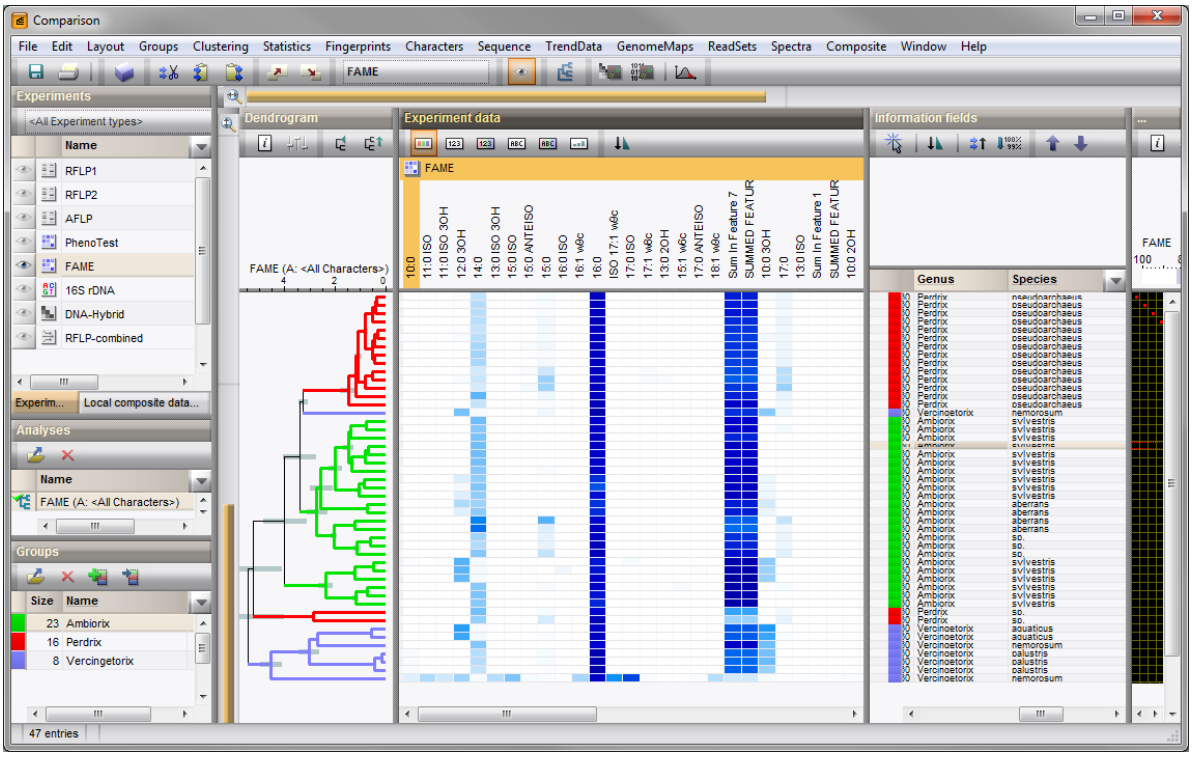

Figure 5: Show group colors on dendrogram.

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.