BIONUMERICS Tutorial:

Clustering fingerprint data

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 fingerprint data. 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  button, located in the toolbar in 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 `https://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 Comparison window

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

2. In the *Database entries* panel of the *Main* window, select all entries except STANDARD: select **Ctrl+A** to select all entries and use the **Ctrl-** key to unselect the entries defined as STANDARD. Alternatively unselect the entries by unchecking the check boxes next to the entries.

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

4. Click on the next to the experiment name **RFLP1** in the *Experiments* panel to display the **RFLP1** patterns in the *Experiment data* panel.

5. Select ***Fingerprints*** > ***Settings*** > ***Show metrics scale*** ( ) to display the metric (e.g. molecular weight) scale of the selected fingerprint type.

6. Press to show the band positions in the *Experiment data* panel.

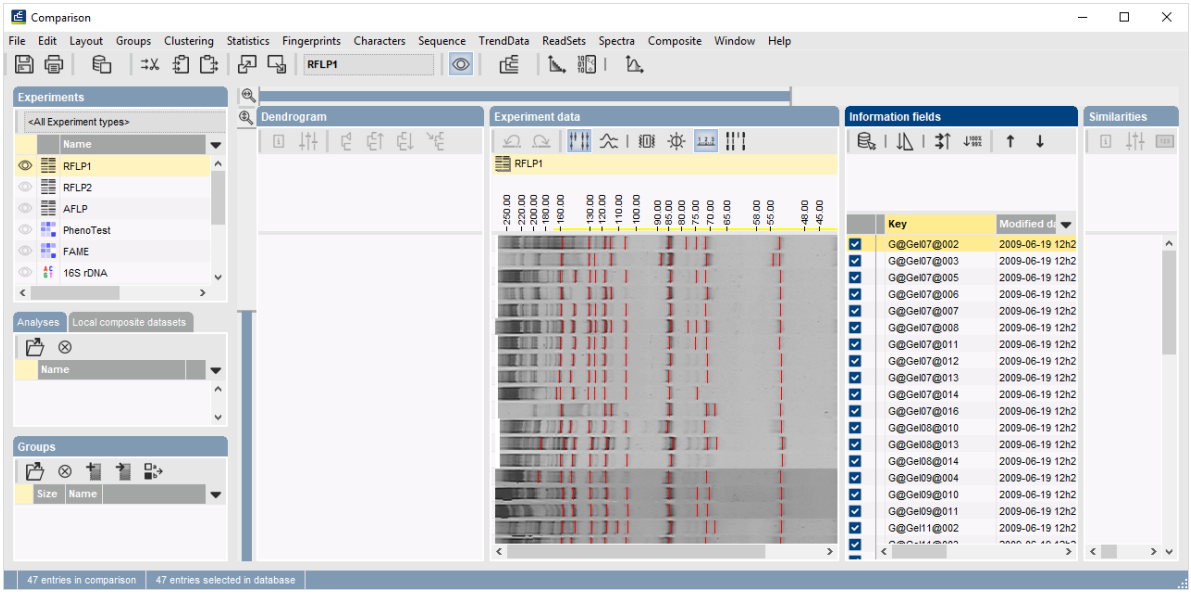

Figure 1: The *Comparison* window.

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 **RFLP1** 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 (see Figure 2).

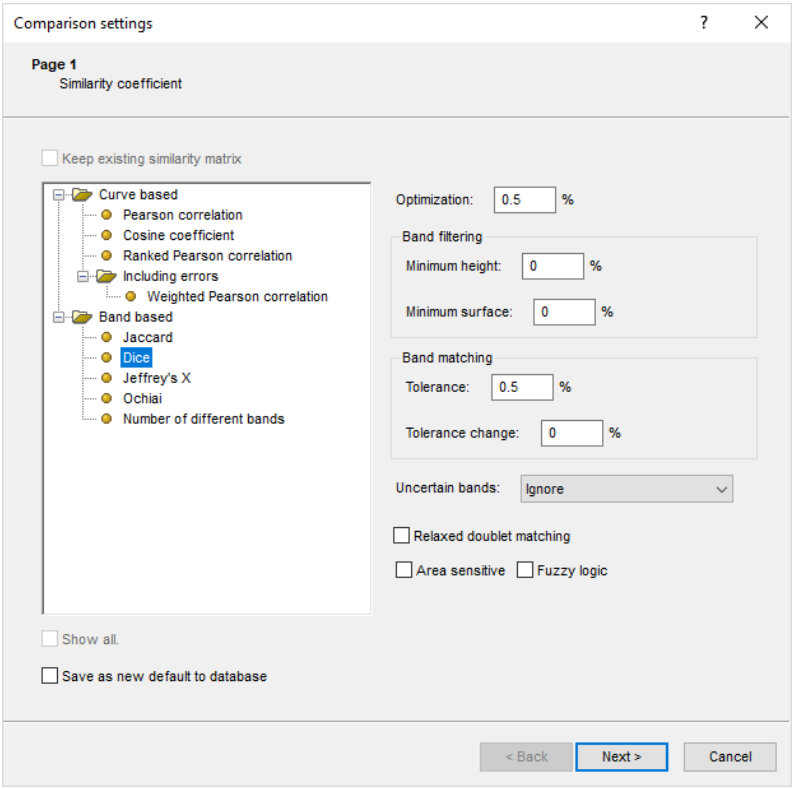

Figure 2: *Similarity coefficient* wizard page: Select similarity coefficient.

In case of fingerprint data, two groups of coefficients can be applied for the calculation of the similarity/distance matrix:

- Curve based coefficients provide similarities based upon densitometric curves
- Band based coefficients measure the similarity based upon common and different bands.

2. Select ***Dice*** from the list.

Additional settings are listed in the right panel.

3. Enter an ***Optimization*** of 0.50%, and a ***Band matching Tolerance*** of 0.50%. Leave all the other settings to 0% (see Figure 2).

The ***Optimization*** setting limits the amount of movement for each fingerprint as a whole. The ***Band matching Tolerance setting*** limits the amount of movement for each band.

4. Press <***Next***>.

In step two the options related to the clustering algorithms are grouped (see Figure 3). 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 will be used.

5. Select ***UPGMA***, check ***Calculate error flags*** and select ***Cophenetic correlation*** from the ***Branch quality*** list (see Figure 3).

If ***Calculate error flags*** is checked, the program will calculate the standard deviations associated with each cluster. The ***Cophenetic Correlation*** is another parameter that expresses the consistency of a cluster. This method calculates the correlation between the dendrogram-derived

similarities and the matrix similarities. The value is calculated for each cluster thus estimating the faithfulness of each sub-cluster of the dendrogram.

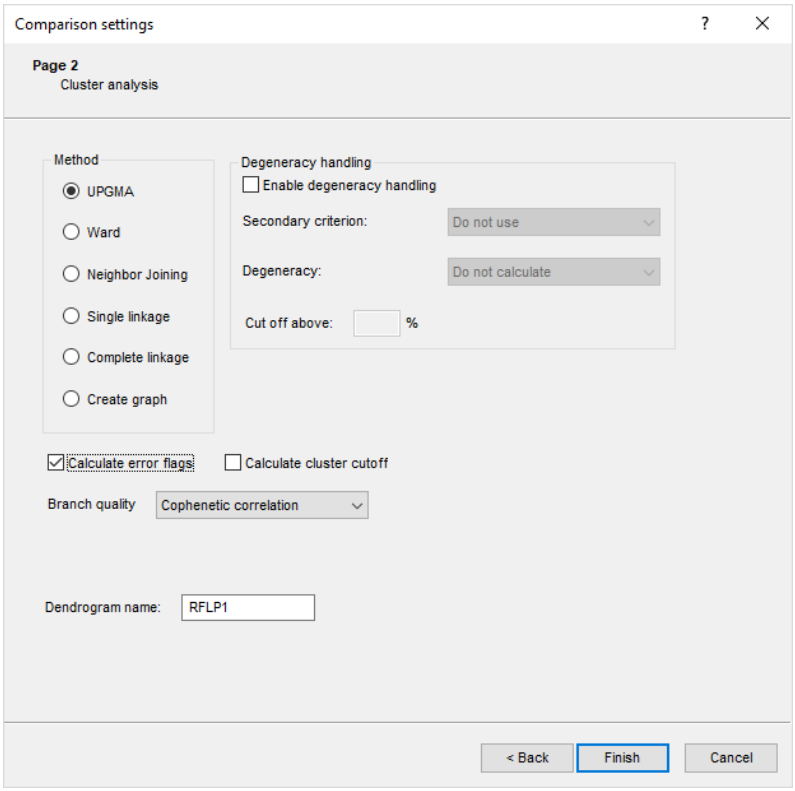

Figure 3: Select clustering algorithm.

6. 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 (see Figure 4).

The ***Cophenetic correlation*** is shown at each branch, together with a colored dot, of which the color ranges between green-yellow-orange-red according to decreasing cophenetic correlation. This makes it easy to detect reliable and unreliable clusters at a glance.

Grey bars are also shown at each node, corresponding to the ***Standard deviation*** of values in that region of the similarity matrix. The average and the standard deviation of similarity values for the selected node are shown above the dendrogram.

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.

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

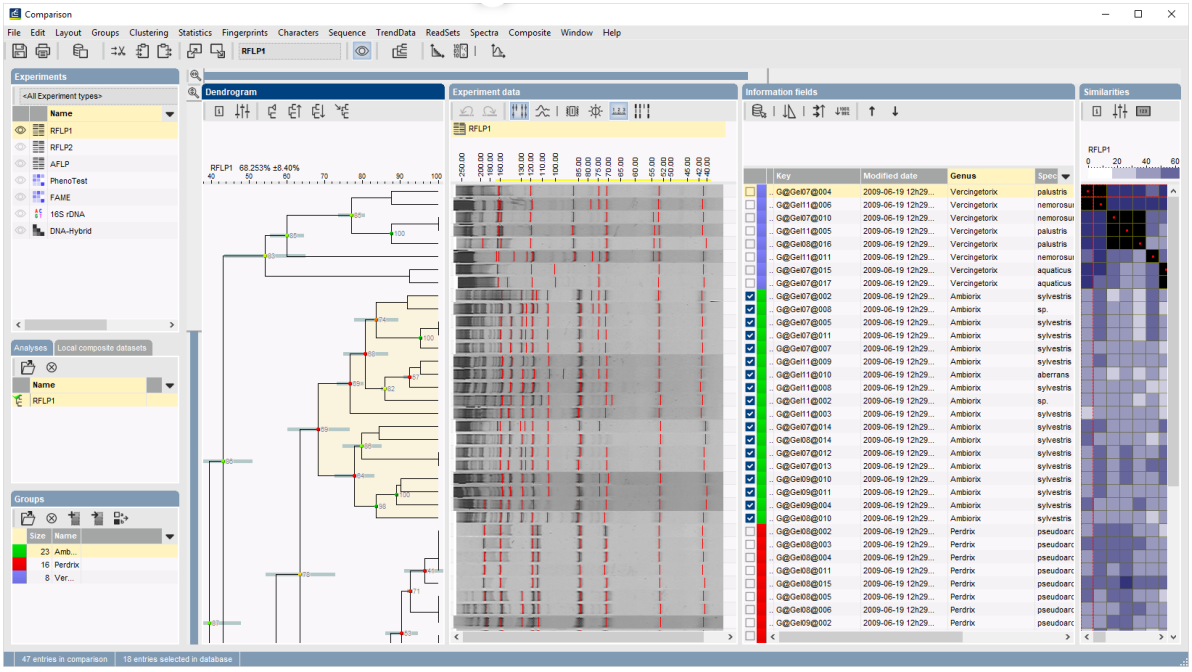

Figure 4: The *Comparison* window with groups defined.

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

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

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

13. 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:

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

15. 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.

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

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

18. Uncheck ***Show error flags***, uncheck ***Show branch quality***, and enable ***Show group colors***. Press <***OK***>.

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

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

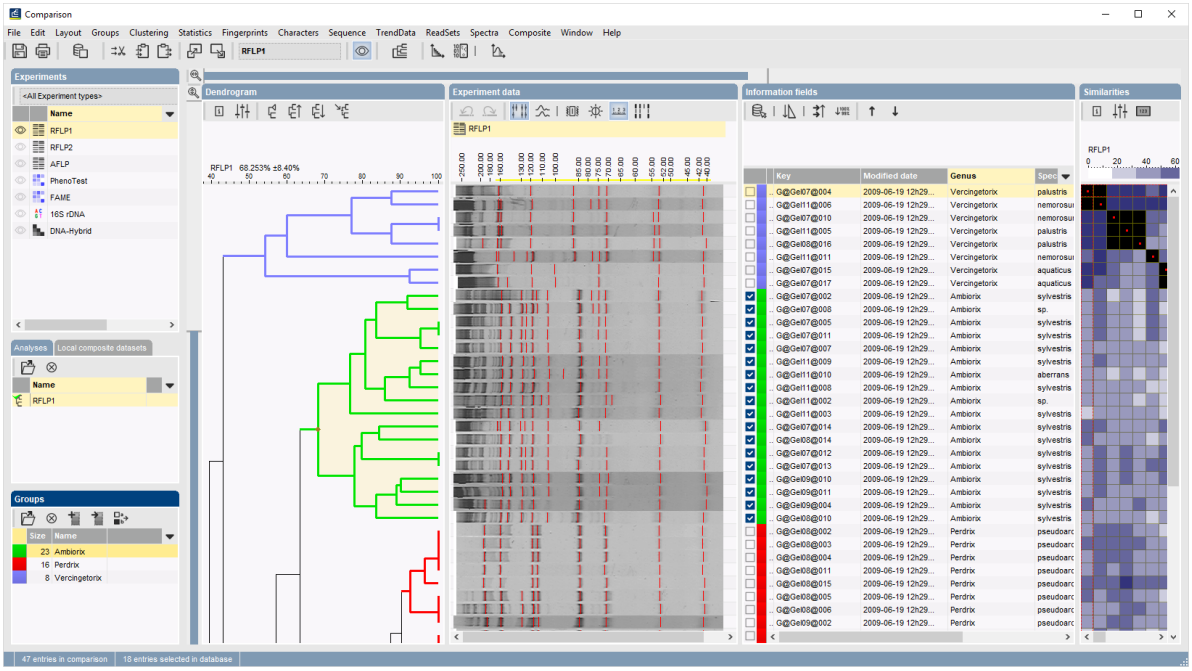

Figure 5: Show group colors on dendrogram.

5 Matrix display functions

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

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

6 Pairwise comparison

1. To view a pairwise comparison between two entries, double-click on the appropriate cell in the matrix (see Figure 6 for an example).
2. When selecting **RFLP1** from the list, the detailed comparison of the band matching is shown in the right panel.
3. Close the *Pairwise comparison* window with ***File*** > ***Exit***.

7 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.

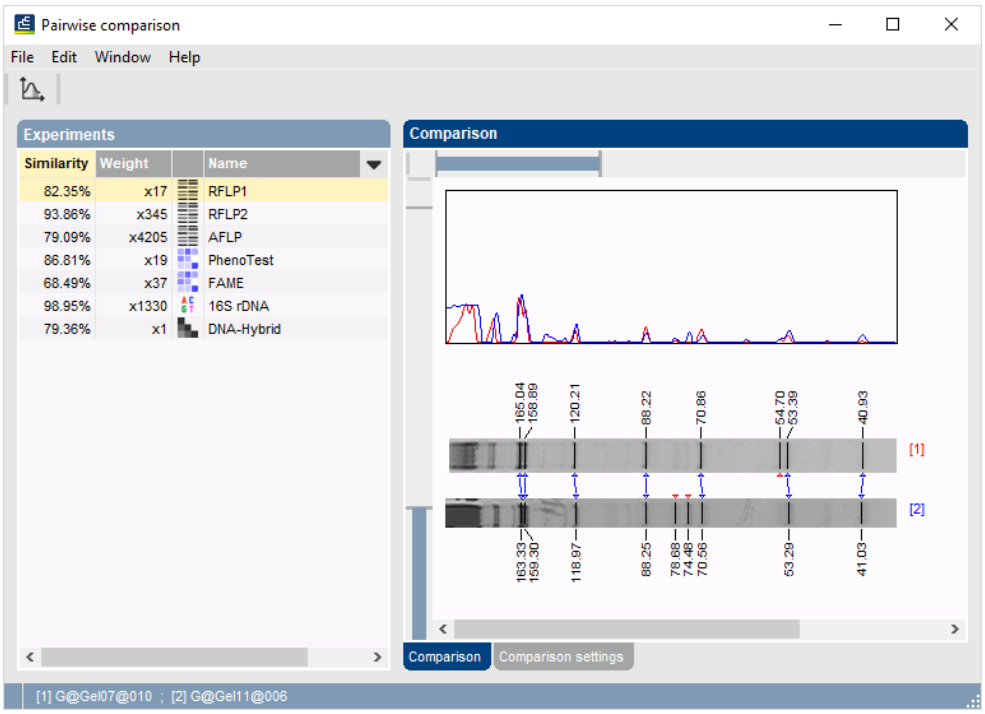

Figure 6: A pairwise comparison.

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

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

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

9. Select ***File*** > ***Exit*** to close the *Comparison print preview* window. Optionally save the print template in the database.