BIONUMERICS Tutorial:

Clustering Omnilog carbon source oxidation 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 trend data. 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 Example data

1. Import the Omnilog .csv trend data files as described in the tutorial: ”Importing carbon source oxidation data from Omnilog csv files”.

Each csv file contains information about the utilization of carbon substrates of a certain strain.

3 Comparison window

1. In the *Database entries* panel of the *Main* window, select all entries in the database for which Omnilog trend curves are present: use the **Ctrl-** key to select the entries, or alternatively right-click on the ***Omnilog*** column in the *Experiment presence* panel and select ***Select entries with experiment***.

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

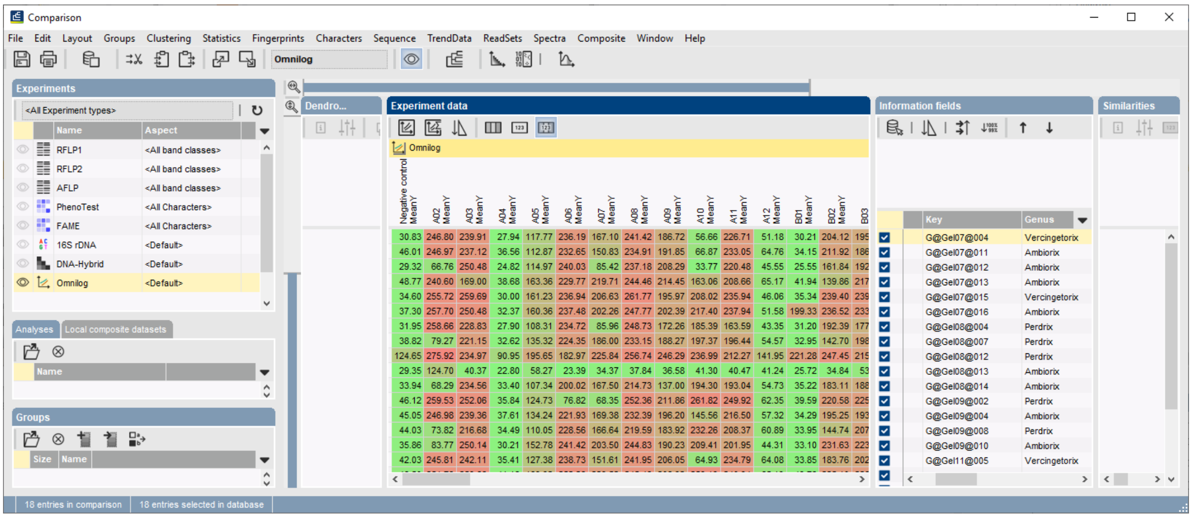
3. Click on the 👁 next to the experiment name **Omnilog** in the *Experiments* panel to display the defined parameter(s) in the *Experiment data* panel (see Figure 1).

4. Select ***TrendData*** > ***Show parameter values colors*** to display the values of the parameter together with the color as defined in the *Trend type* window.

5. Select a parameter in the *Experiment data* panel and select ***TrendData*** > ***Sort entries by parameter value*** (⇅).

The entries are sorted according to increasing value of the selected parameter.

6. A tab-delimited text file of the entries and trend data values contained in the comparison can be exported with ***TrendData*** > ***Export character table***.



Name	Aspect
RFLP1	<All band classes>
RFLP2	<All band classes>
AFLP	<All band classes>
PhenoTest	<All Characters>
FAME	<All Characters>
16S rDNA	<Default>
DNA-Hybrid	<Default>
Omnilog	<Default>

Negative control MeanY	A02 MeanY	A03 MeanY	A04 MeanY	A05 MeanY	A06 MeanY	A07 MeanY	A08 MeanY	A09 MeanY	A10 MeanY	A11 MeanY	A12 MeanY	B01 MeanY	B02 MeanY	B03
30.83	246.80	239.91	27.94	117.77	236.19	167.10	241.42	186.72	56.66	226.71	51.18	30.21	204.12	195
46.01	246.97	237.12	36.56	112.87	232.65	150.83	234.91	191.85	66.87	233.05	64.76	34.15	211.92	186
29.32	66.76	250.48	24.82	114.97	240.03	85.42	237.18	208.29	33.77	220.48	45.55	25.55	161.84	192
48.77	240.60	169.00	38.68	163.36	229.77	219.71	244.46	214.45	163.06	208.66	65.17	41.94	139.86	217
34.60	255.72	259.69	30.00	161.23	236.94	206.63	261.77	195.97	208.02	235.94	46.06	35.34	239.40	239
37.30	257.70	250.48	32.37	160.36	237.48	202.26	247.77	202.39	217.40	237.94	51.58	199.33	236.52	233
31.95	258.66	228.83	27.90	108.31	234.72	85.96	248.73	172.26	185.39	163.59	43.35	31.20	192.39	177
38.82	79.27	221.15	32.62	135.32	224.35	186.00	233.15	188.27	197.37	196.44	54.57	32.95	142.70	198
124.65	275.92	234.97	90.95	195.65	182.97	225.84	256.74	246.29	236.99	212.27	141.95	221.28	247.45	215
29.35	124.70	40.37	22.80	58.27	23.39	34.37	37.84	36.58	41.30	40.47	41.24	25.72	34.84	53
33.94	68.29	234.56	33.40	107.34	200.02	167.50	214.73	137.00	194.30	193.04	54.73	35.22	183.11	188
46.12	259.53	252.06	35.84	124.73	76.82	68.35	252.36	211.86	261.82	249.92	62.35	39.59	220.58	225
45.05	246.98	239.36	37.61	134.24	221.93	169.38	232.39	196.20	145.56	216.50	57.32	34.29	195.25	193
44.03	73.82	216.68	34.49	110.05	228.56	166.64	219.59	183.92	232.26	208.37	60.89	33.95	144.74	207
35.86	83.77	250.14	30.21	152.78	241.42	203.50	244.83	190.23	209.41	201.95	44.31	33.10	231.63	223
42.03	245.81	242.11	35.41	127.38	238.73	151.61	241.95	206.05	64.93	234.79	64.08	33.85	183.76	202

	Key	Genus
☑	G@Gel07@004	Vercingetorix
☑	G@Gel07@011	Ambiorix
☑	G@Gel07@012	Ambiorix
☑	G@Gel07@013	Ambiorix
☑	G@Gel07@015	Vercingetorix
☑	G@Gel07@016	Ambiorix
☑	G@Gel08@004	Perdrix
☑	G@Gel08@007	Perdrix
☑	G@Gel08@012	Perdrix
☑	G@Gel08@013	Ambiorix
☑	G@Gel08@014	Ambiorix
☑	G@Gel09@002	Perdrix
☑	G@Gel09@004	Ambiorix
☑	G@Gel09@008	Perdrix
☑	G@Gel09@010	Ambiorix
☑	G@Gel11@005	Vercingetorix

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 **Omnilog** 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).

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

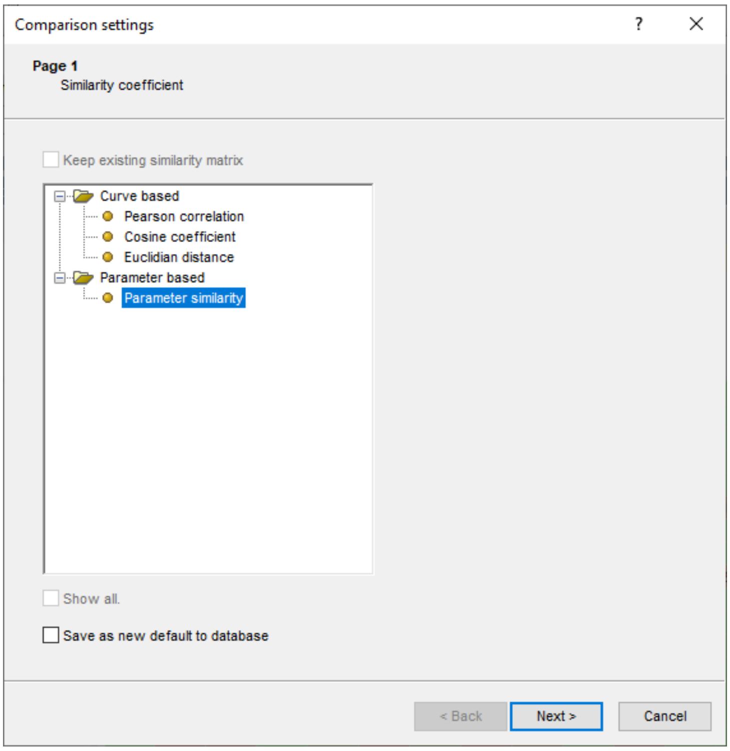
- Curve based coefficients: provide similarities based upon the original data points of the curves.
- Parameter based coefficient: measures the similarity by comparing the values of the parameter(s), defined in the *Trend type* window.

2. Select a coefficient from the list, e.g. ***Parameter similarity*** 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 will be used.

3. Select ***UPGMA*** and select ***Cophenetic correlation*** from the ***Branch quality*** list (see Figure 3).

The ***Cophenetic Correlation*** is a 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 subcluster of the dendrogram.



Comparison settings

Page 1
Similarity coefficient

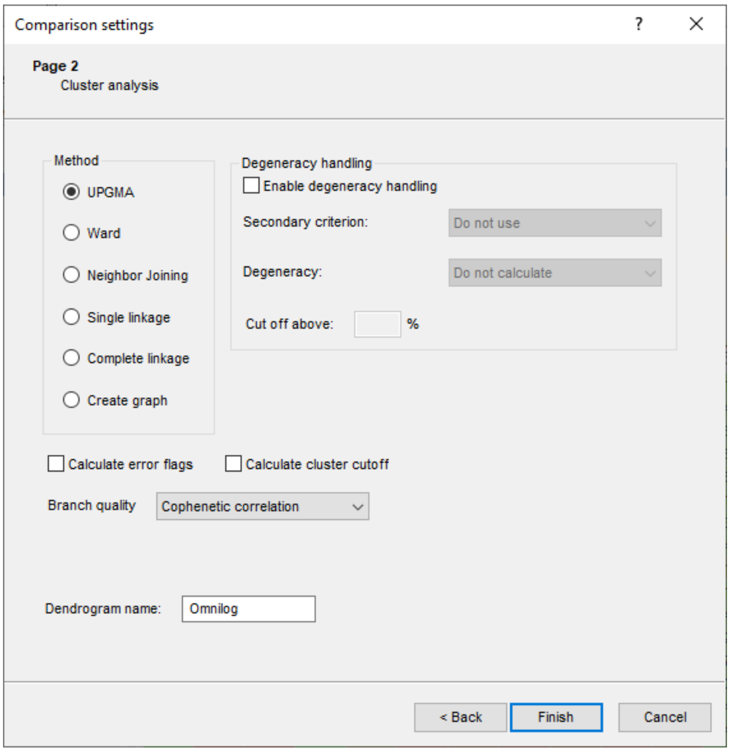
□ Keep existing similarity matrix

- Curve based
  - Pearson correlation
  - Cosine coefficient
  - Euclidian distance
- Parameter based
  - Parameter similarity

□ Show all.

□ Save as new default to database

Figure 2: Select similarity coefficient.



Comparison settings

Page 2
Cluster analysis

Method
● UPGMA
○ Ward
○ Neighbor Joining
○ Single linkage
○ Complete linkage
○ Create graph

Degeneracy handling
□ Enable degeneracy handling
Secondary criterion: Do not use
Degeneracy: Do not calculate
Cut off above: %

□ Calculate error flags □ Calculate cluster cutoff

Branch quality: Cophenetic correlation

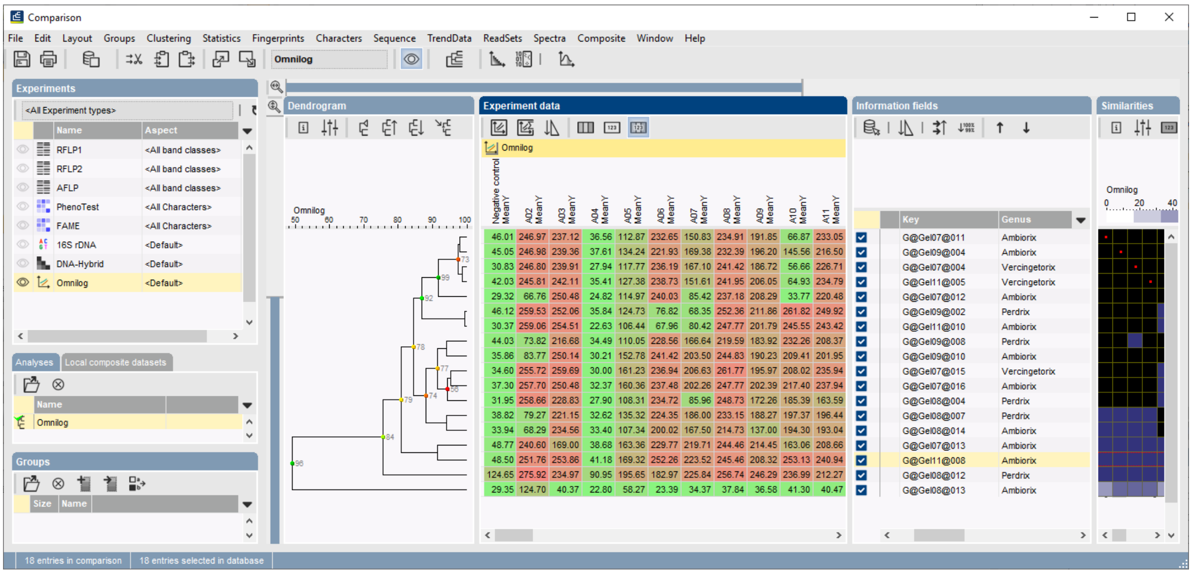
Dendrogram name: Omnilog

Figure 3: Select clustering algorithm.

4. 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).



Name	Aspect
RFLP1	<All band classes>
RFLP2	<All band classes>
AFLP	<All band classes>
PhenoTest	<All Characters>
FAME	<All Characters>
16S rDNA	<Default>
DNA-Hybrid	<Default>
Omnilog	<Default>

Name	
Omnilog	

Omnilog

Negative control MeanY	A02 MeanY	A03 MeanY	A04 MeanY	A05 MeanY	A06 MeanY	A07 MeanY	A08 MeanY	A09 MeanY	A10 MeanY	A11 MeanY		Key	Genus
46.01	246.97	237.12	36.56	112.87	232.65	150.83	234.91	191.85	66.87	233.05	☑	G@Gel07@011	Ambiorix
45.05	246.98	239.36	37.61	134.24	221.93	169.38	232.39	196.20	145.56	216.50	☑	G@Gel09@004	Ambiorix
30.83	246.80	239.91	27.94	117.77	236.19	167.10	241.42	186.72	56.66	226.71	☑	G@Gel07@004	Vercingetorix
42.03	245.81	242.11	35.41	127.38	238.73	151.61	241.95	206.05	64.93	234.79	☑	G@Gel11@005	Vercingetorix
29.32	66.76	250.48	24.82	114.97	240.03	85.42	237.18	208.29	33.77	220.48	☑	G@Gel07@012	Ambiorix
46.12	259.53	252.06	35.84	124.73	76.82	68.35	252.36	211.86	261.82	249.92	☑	G@Gel09@002	Perdrix
30.37	259.06	254.51	22.63	106.44	67.96	80.42	247.77	201.79	245.55	243.42	☑	G@Gel11@010	Ambiorix
44.03	73.82	216.68	34.49	110.05	228.56	166.64	219.59	183.92	232.26	208.37	☑	G@Gel09@008	Perdrix
35.86	83.77	250.14	30.21	152.78	241.42	203.50	244.83	190.23	209.41	201.95	☑	G@Gel09@010	Ambiorix
34.60	255.72	259.69	30.00	161.23	236.94	206.63	261.77	195.97	208.02	235.94	☑	G@Gel07@015	Vercingetorix
37.30	257.70	250.48	32.37	160.36	237.48	202.26	247.77	202.39	217.40	237.94	☑	G@Gel07@016	Ambiorix
31.95	258.66	228.83	27.90	108.31	234.72	85.96	248.73	172.26	185.39	163.59	☑	G@Gel08@004	Perdrix
38.82	79.27	221.15	32.62	135.32	224.35	186.00	233.15	188.27	197.37	196.44	☑	G@Gel08@007	Perdrix
33.94	68.29	234.56	33.40	107.34	200.02	167.50	214.73	137.00	194.30	193.04	☑	G@Gel08@014	Ambiorix
48.77	240.60	169.00	38.68	163.36	229.77	219.71	244.46	214.45	163.06	208.66	☑	G@Gel07@013	Ambiorix
48.50	251.76	253.86	41.18	169.32	252.26	223.52	245.46	208.32	253.13	240.94	☑	G@Gel11@008	Ambiorix
124.65	275.92	234.97	90.95	195.65	182.97	225.84	256.74	246.29	236.99	212.27	☑	G@Gel08@012	Perdrix
29.35	124.70	40.37	22.80	58.27	23.39	34.37	37.84	36.58	41.30	40.47	☑	G@Gel08@013	Ambiorix

18 entries in comparison | 18 entries selected in database

Figure 4: The *Comparison* window.

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.

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

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

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

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

9. 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:

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

11. 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.

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

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

14. Uncheck ***Show branch quality*** and press <***OK***> to remove the cophenetic correlation from the tree.

15. Select ***Clustering*** > ***Show information*** (ⓘ) to display a report containing the comparison settings. Close the report.

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*** (123).

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

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*** (ᴀA) or ***Layout*** > ***Reduce image size*** (Aᴀ).

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.