BIONUMERICS Tutorial:

Follow-up analysis of MLST data

1 Aim

In this tutorial we will perform some cluster analyses on MLST data. We will also see how we can alter the layout of the clusterings and how to export the pictures to use it in a publication, presentation, etc.

2 Preparing the database

2.1 Introduction to the MLST demo database

The **MLST demo database** contains for 500 *Neisseria meningitidis* isolates following information: a unique identifier (”Key”), a strain number, an MLST sequence type that was deduced from the analysis (”ST”), the clonal complex information (”CC”), the serogroup, the country where the strains originate from, the year of isolation, the species and the disease in which the strains were involved (see Figure 1).

The allele number is reported for each of the seven loci sequenced (sequence types **abcZ**, **adk**, **aroE**, **fumC**, **gdh**, **pdhC** and **pgm**) for all 500 strains and is stored in the **MLST** character type experiment.

The **MLST demo database** can be downloaded directly from the *BIONUMERICS Startup* window (see 2.2), or the data can be imported from a file available on our website, in a new, empty BIONUMERICS database (see 2.3), or the database can be restored from a back-up file available on our website (see 2.4).

2.2 Option 1: Download the demo database from the Startup Screen

1. Click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 2).

2. Select the **Neisseria MLST demo database** from the list and select ***Database*** > ***Download*** ().

3. Confirm the installation of the database and press <***OK***> after successful installation of the database.

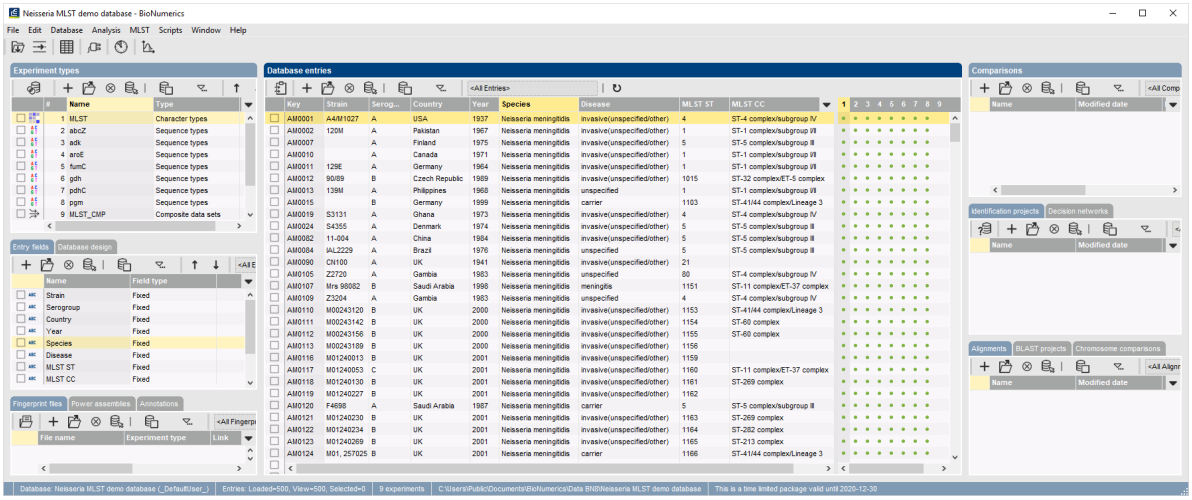

Figure 1: The *Main* window of the MLST demo database.

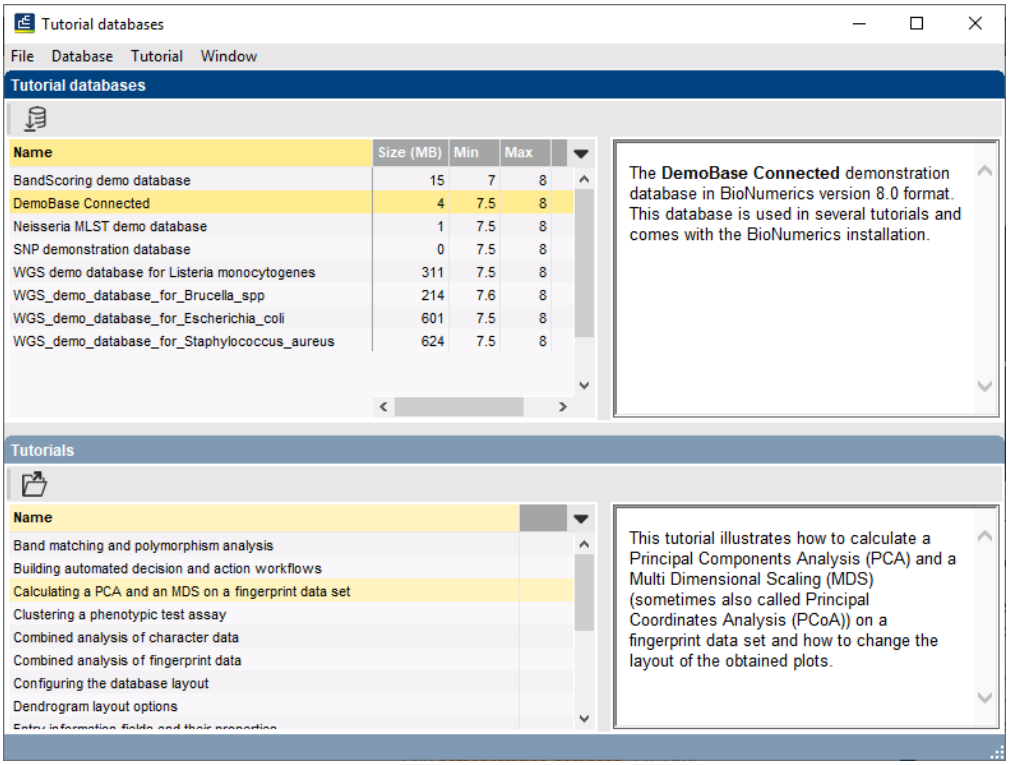

Figure 2: The *Tutorial databases* window.

4. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **Neisseria MLST demo database** appears in the *BIONUMERICS Startup* window.

5. Double-click the **Neisseria MLST demo database** in the *BIONUMERICS Startup* window to open the database.

The *Main* window should look like Figure 1.

2.3 Option 2: Import the data from an Excel file in a new database

6. Create a new database or open an existing database.

7. Import the MLST dataset from the example Excel file `Neisseria MLST.xlsx` as described in the tutorial: ”Importing MLST data from an Excel file”. The Excel file contains preprocessed MLST information for about 500 *Neisseria meningitidis* strains.

After import the *Main* window should look like Figure 1, but without the sequence type experiments.

2.4 Option 3: Restore demo database from back-up file

A BIONUMERICS back-up file of the **Neisseria MLST demo database** is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

8. Download the file `Neisseria.bnbk` from https://www.applied-maths.com/download/sample-data, under ’Neisseria MLST demo database’.

In contrast to other browsers, some versions of Internet Explorer rename the `Neisseria.bnbk` database backup file into `Neisseria.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.

9. In the *BIONUMERICS Startup* window, press the  button. From the menu that appears, select ***Restore database...***.

10. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.

11. Specify a new name for this demonstration database, e.g. “Neisseria MLST demo database”.

12. Click <***OK***> to start restoring the database from the backup file.

13. Once the process is complete, click <***Yes***> to open the database.

The *Main* window should look like Figure 1.

3 Working in the database

To visually discriminate different information field *states* in the database or in comparisons, colors can be assigned via the information field properties. For example, to assign different colors to the three serogroups (A, B and C) in the database, proceed as follows:

1. Right-click on the **Serogroup** information field in the *Main* window and choose ***Field properties*** from the floating menu. Alternatively, double-click on **Serogroup** in the *Entry fields* panel.

2. In the *Database field properties* dialog box, press <***Add all***> to create all existing states for the **Serogroup** field. Confirm the action.

All information field states that are present in the database (A, B, C), appear in the ***Field states*** list on the left.

3. Check ***Use colors*** to display a specific color code for each field state and press <***OK***>.

4. Repeat the above instructions for the information field **Country** (see Figure 3).

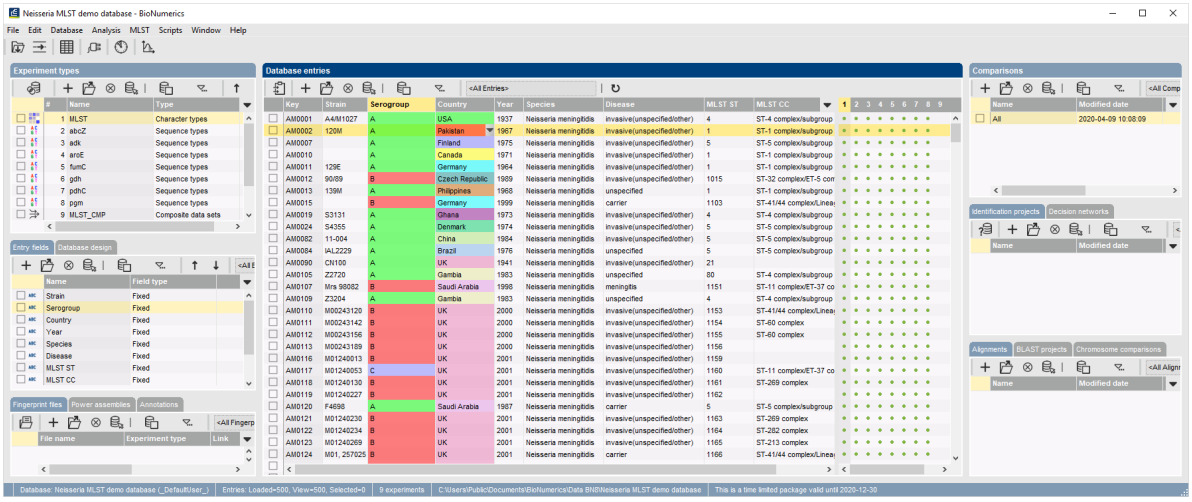

Figure 3: The *Main* window.

4 Creating a comparison

The MLST allele IDs for all seven housekeeping genes are stored in the character type **MLST**, and the individual housekeeping gene sequences are stored in the corresponding sequence type experiment.

1. Click on a green colored dot in the *Experiment presence* panel for experiment **MLST** to open the experiment card for an entry.

The allele numbers are displayed in the experiment card next to the corresponding housekeeping gene names (see Figure 4).

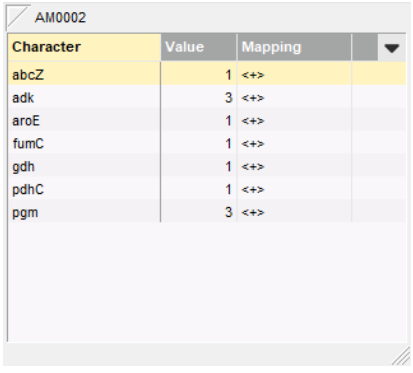

Figure 4: Character experiment card.

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

3. In the *Database entries* panel of the *Main* window, select all entries using ***Edit*** > ***Select all*** (**Ctrl+A**).

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

5. Click on the next to the experiment name **MLST** in the *Experiments* panel and select ***Characters*** > ***Show values*** ( ) to display the allele numbers in the *Experiment data* panel.

6. In the *Comparison* window, right-click in the header of the "CC" field and select ***Create groups from database field*** from the floating menu. Alternatively select ***Groups*** > ***Create groups from database field***.

7. In the *Group creation preferences* dialog box, make sure ***Create largest groups first*** is selected, select ***Skip empty content***, specify a maximum count of ***20*** and press <***OK***> twice.

Every clonal complex with at least three members is now assigned to a unique group. The 20 groups appear in the *Groups* panel along with their color, size and name (see Figure 5).

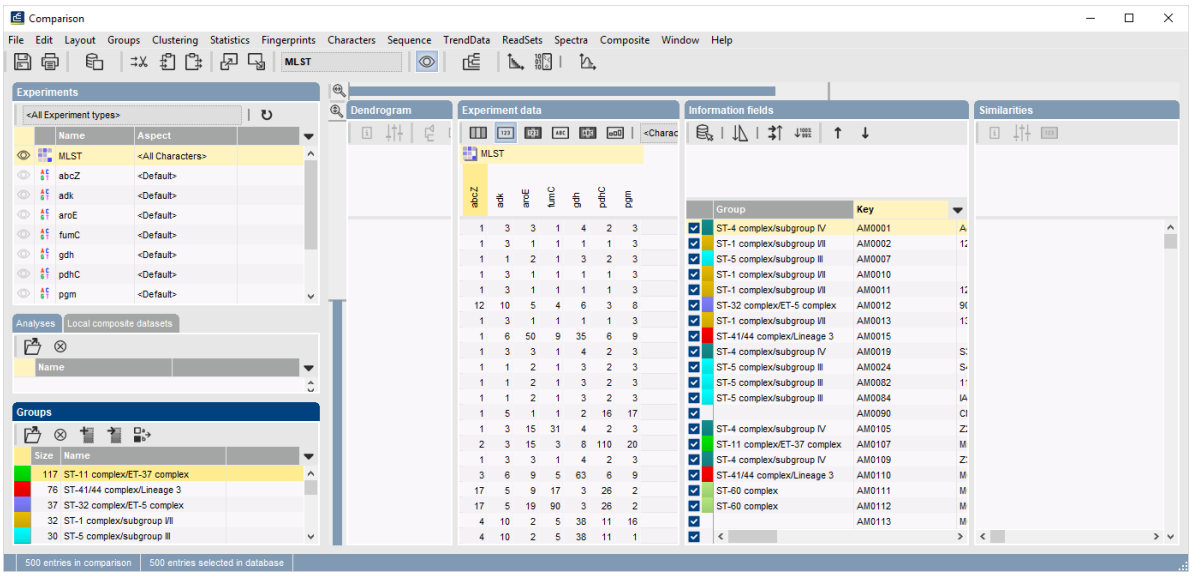

Figure 5: The *Comparison* window with comparison groups defined.

5 Creating a similarity based clustering

1. Make sure **MLST** 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. Due to the arbitrariness of the allele numbers, the similarity coefficient for clustering MLST data is the categorical coefficient. The categorical coefficient compares the allele numbers to see if they are the same or different but does not quantify the difference.

2. Select ***Categorical (values)*** 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: "MLST(<All characters>)") will be used.

3. Select ***UPGMA***, change the name of the analysis (e.g. **MLST 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.

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

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

The dendrogram branches are now colored according to the group colors (see Figure 6).

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

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

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

16. Select ***Categorical (values)*** from the list and press <***Next***>.

17. Select ***Single linkage***, change the name of the analysis (e.g. **MLST Single linkage**) and <***Finish***> to start the cluster analysis.

Both analyses are now listed in the *Analyses* panel. Switching between the different dendrograms can be done by simply double-clicking on the analysis name.

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

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

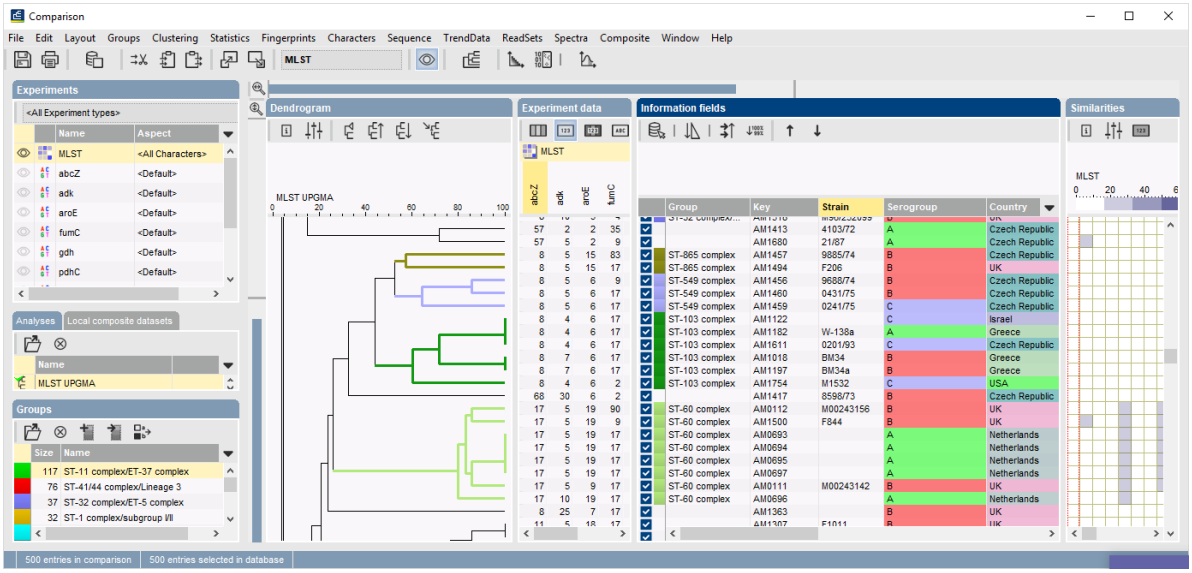

Figure 6: The *Comparison* window: UPGMA dendrogram.

The *Comparison print preview* window now appears.

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

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

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

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

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

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

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

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

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

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

29. Close the saved comparison with ***File*** > ***Exit***.

6 Creating a minimum spanning tree

A minimum spanning tree in BIONUMERICS is calculated in the *Advanced cluster analysis* window. This window can be launched from the *Comparison* window.

1. Double-click on the saved comparison **All** in the *Comparisons* panel in the *Main* window.

2. Make sure ***MLST*** is selected in the *Experiments* panel of the *Comparison* window.

3. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** or press the button and select ***Advanced cluster analysis*** to launch the *Create network* wizard.

The predefined template ***MST for categorical data*** uses the categorical coefficient for the calculation of the similarity matrix, and will calculate a standard minimum spanning tree with single and double locus variance priority rules.

4. Specify an analysis name (for example **MLST MST**), make sure ***MLST*** is selected, select ***MST for categorical data***, and press <***Next***>.

To view and modify the settings of a selected template check the option ***Modify template settings for new analysis***.

The *Advanced cluster analysis* window pops up. The *Network panel* displays the minimum spanning tree, the upper right panel (*Entry list*) displays the entries that are present in the tree. The *Cluster analysis method panel* displays the settings used, in this example the priority rules that result in the displayed network.

The colors of the comparison groups are automatically shown as node colors, but this can very easily be changed to a field state grouping defined in the *Main* window:

5. Press or choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box.

6. In the *Node colors tab* select the **Serogroup** from the list and press <***OK***>.

The node colors are updated according to the serogroups.

7. A node or branch can be selected by clicking on them. To select several nodes/branches hold the **Shift**-key.

8. The zoom slider on the left always further zooming in or out on the network. The zoom slider on top adjusts the size of the nodes.

9. Select ***Display*** > ***Zoom to fit*** or press to optimize the view of the tree.

10. Press or choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box again.

11. In the *Branch labels and sizes tab*, check ***Use logarithmic scaling***.

12. In the *Node colors tab* select the **Comparison groups** option again from the list and make sure the option ***Separate entries*** is unchecked.

13. Press <***OK***> to apply the new settings (see Figure 7).

In the *Advanced cluster analysis* window it is possible to create *partitions*. In case of an MST, the partitioning algorithm will group nodes in partitions (complexes) when the distance between the connected nodes is less than or equal to a distance entered by the user. As soon as a connection has a longer distance, the partition ends.

14. A partitioning can be created with ***Edit*** > ***Create partitioning*** or using the button. This calls the *Partitioning* dialog box.

15. For the current example, enter a ***Maximum distance between nodes in the same partition*** of 2 and a ***Minimum number of entries in a partition*** of 2. Choose ***Color from majority*** and press <***OK***>.

The color of the partitions is adopted from the node colors. In case the nodes have different colors, the color from the majority is taken.

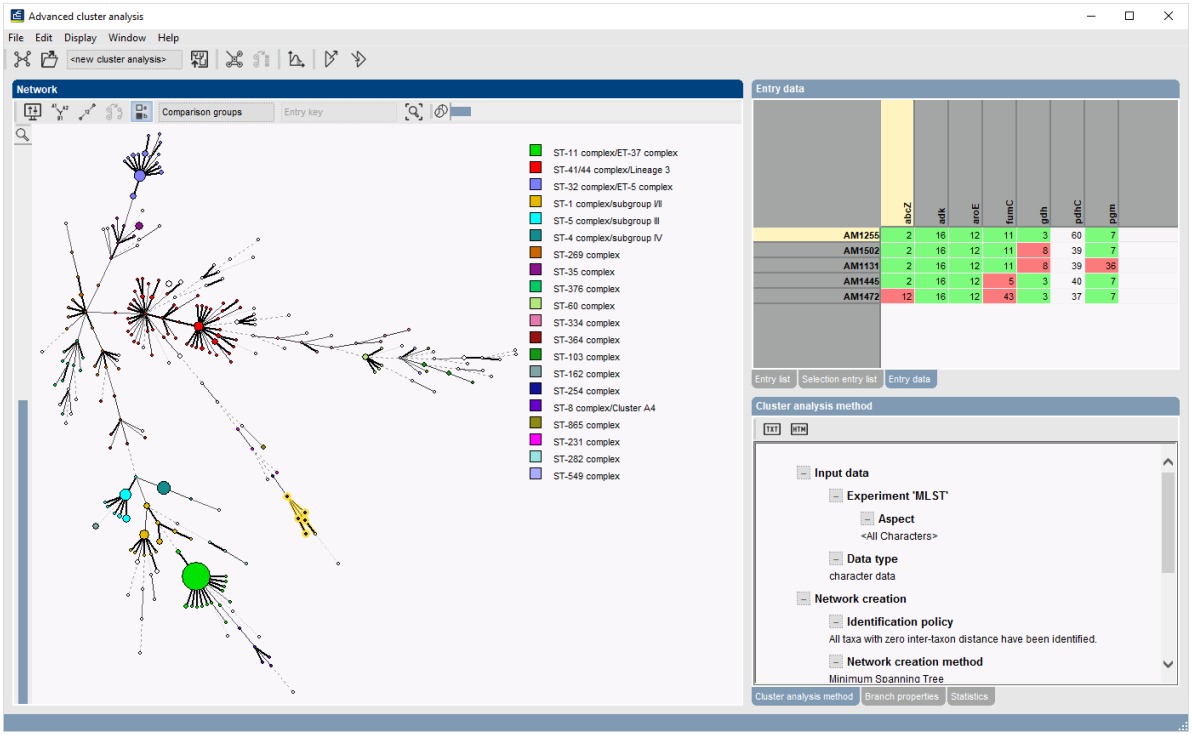

Figure 7: The *Advanced cluster analysis* window: minimum spanning tree.

From this picture it is clear that the definition of a partitioning in an MST corresponds to the clonal complexes as defined for MLST and similar allele-based typing techniques.

16. The image can be exported with ***File*** > ***Export image***.

17. Close the *Advanced cluster analysis* window with ***File*** > ***Exit***.

18. Save the comparison with ***File*** > ***Save*** (, **Ctrl+S**) and close the comparison.

7 Creating a maximum parsimony tree

The **MLST_CMP** experiment, used in this section is only available when the **MLST demo database** was downloaded directly from the *BIONUMERICS Startup* window (see 2.2), or when the database was restored from the back-up file available on our website (see 2.4).

1. Double-click on the **MLST_CMP** experiment in the *Experiment types* panel of the *Main* window to call the *Composite data type* window.

All seven housekeeping gene experiments are activated in the composite data set (see Figure 8).

2. Close the *Composite data type* window with ***File*** > ***Exit***.

3. Double-click on any of the seven sequence type experiments in the *Experiment types* panel of the *Main* window and select ***Settings*** > ***Character conversion settings...*** ().

Based on the settings of each housekeeping gene experiment, only the mutated positions will be retained in the composite data set (see Figure 9).

4. Close the *Character conversion settings* dialog box and *Sequence type* window with ***File*** > ***Exit***.

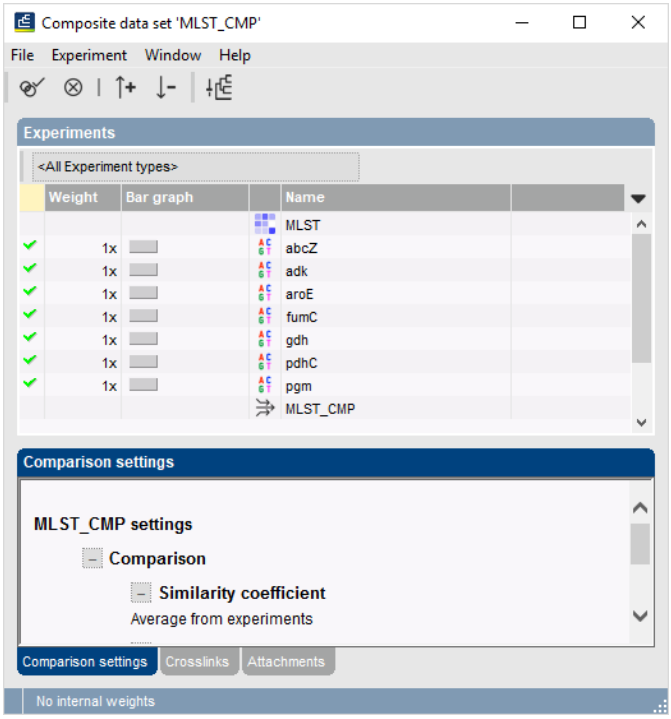

Figure 8: Composite data set.

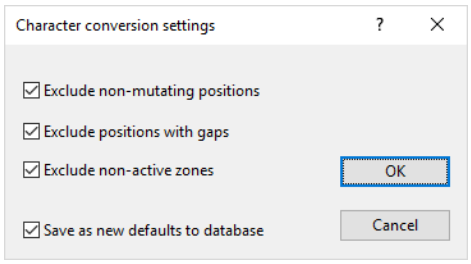

Figure 9: Character conversion settings.

5. Double-click on the saved comparison **All** in the *Comparisons* panel in the *Main* window.

6. Click on the next to the experiment name **MLST_CMP** in the *Experiments* panel to display all mutating positions in the *Experiment data* panel.

Just like a minimum spanning tree, a maximum parsimony tree in BIONUMERICS is calculated in the *Advanced cluster analysis* window.

7. Make sure ***MLST_CMP*** is selected in the *Experiments* panel of the *Comparison* window.

8. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** or press the button and select ***Advanced cluster analysis*** to launch the *Create network* wizard.

The predefined template ***Maximum parsimony tree*** treats the mutations as categorical data, and will calculate a standard maximum parsimony tree.

9. Specify an analysis name (for example **MLST Max parsimony**), make sure ***MLST_CMP*** is selected, select ***Maximum parsimony tree***, and press <***Next***>.

To view and modify the settings of a selected template check the option ***Modify template settings for new analysis***.

The *Advanced cluster analysis* window pops up (see Figure 10). The *Network panel* displays the maximum parsimony tree.

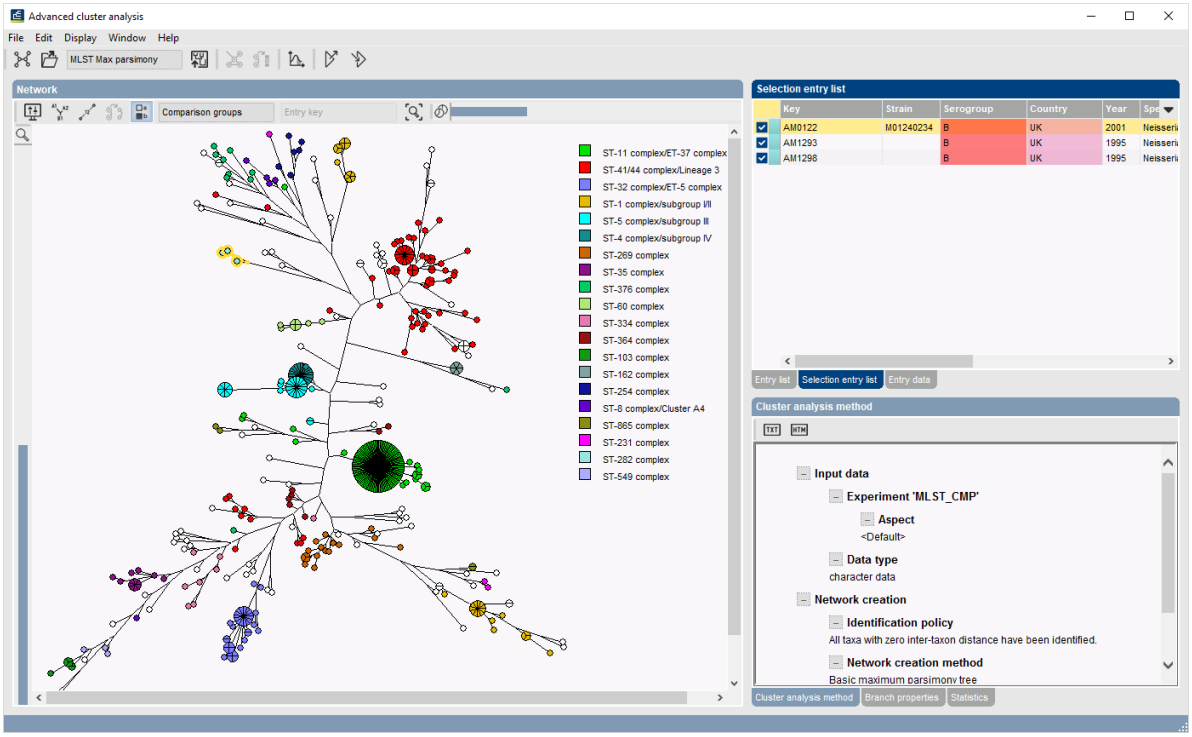

Figure 10: The *Advanced cluster analysis* window.

10. Close the *Advanced cluster analysis* window and *Comparison* window with ***File*** > ***Exit***.