BioNumerics Tutorial:

Creating a minimum spanning tree based on MLST data

1 Aim

In this tutorial we will create a minimum spanning tree based on MLST data. We will also see how we can alter the layout of the minimum spanning tree and how to export the picture 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).

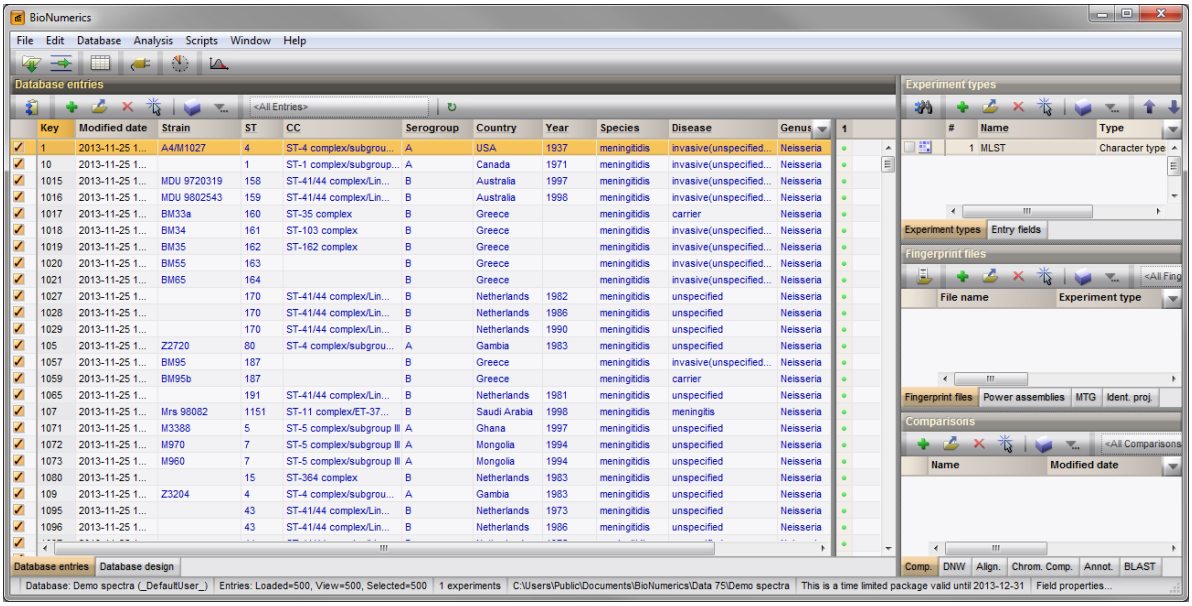

Figure 1: The *Main* window of the MLST demo database.

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

1. Click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window.

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

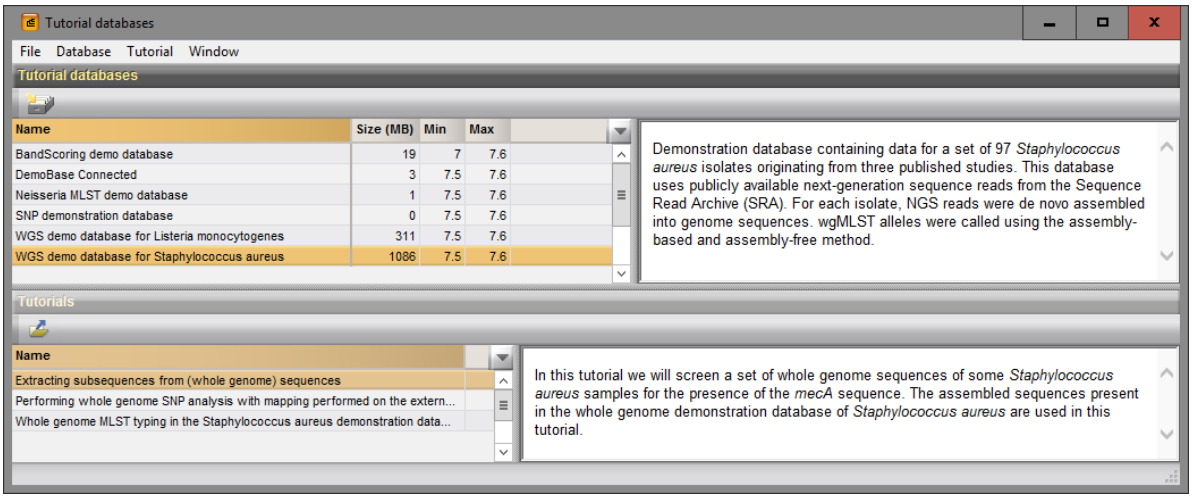

Figure 2: The *Tutorial databases* window, used to download the Neisseria MLST demonstration database.

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

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

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

The **Neisseria demo database** appears in the *BioNumerics Startup* window.

5. Double-click the **Neisseria 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.

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 `http://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*** remains 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

The character data is stored in the character type **MLST**.

1. To view the values in a list, double-click on the experiment **MLST** in the *Experiment types* panel, select ***Settings*** > ***General settings...*** (), select the *Experiment card tab* and change the representation to ***List***. Close the two windows.
2. Click on a green colored dot in the *Experiment presence* panel to open the experiment card for an entry.

The imported allele numbers are displayed in the experiment card next to the corresponding housekeeping gene names.

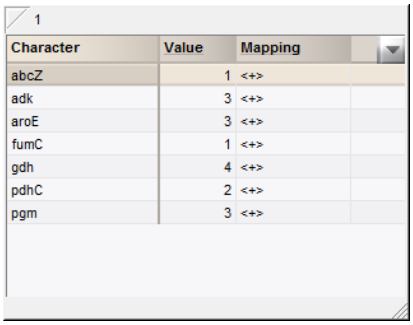

Figure 3: The experiment card.

3. Close the experiment card by clicking in the left upper corner of the card.
4. Right-click on the **Serogroup** information field in the *Main* window and choose ***Field properties*** from the floating menu.
5. Press <***Add all***> to create all existing states for the **Serogroup** field. Confirm the action.
6. Check ***Use colors*** to display a specific color code for each field state (see Figure 5).

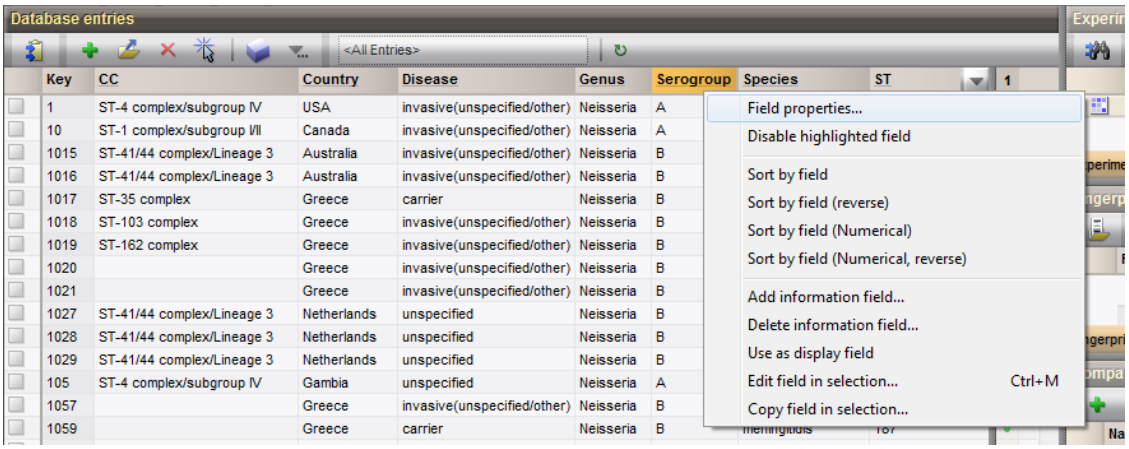

Figure 4: Field properties.

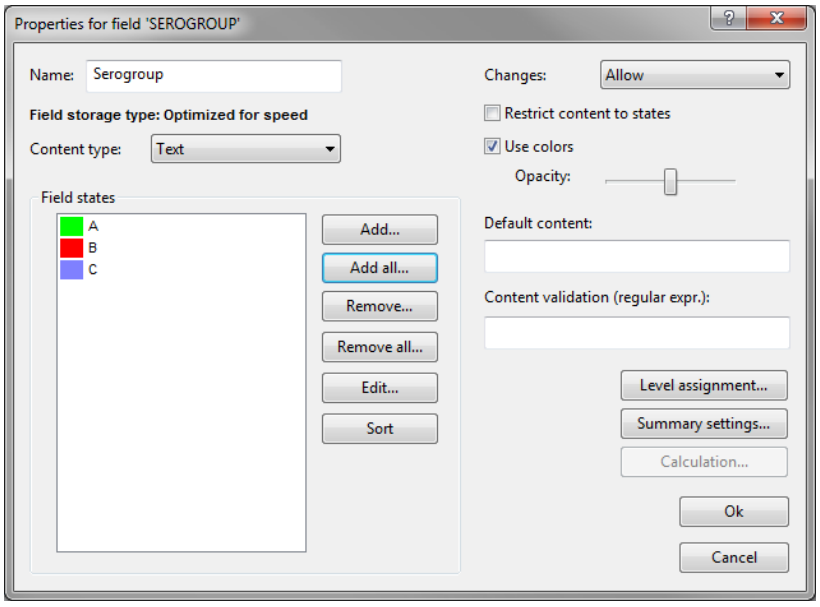

Figure 5: The *Database field properties* dialog box.

7. Press <***OK***> to accept the new settings.

The *Database entries* panel is updated (see Figure 6).

Since it is also possible to create groups based on the field content in the *Comparison* window, we will use the content of the **CC** column as an example there (see 4).

4 Comparison window

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

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 **MLST** in the *Experiments* panel and select ***Characters*** > ***Show values*** ( ) to display the allele numbers in the *Experiment data* panel (see Figure 7).

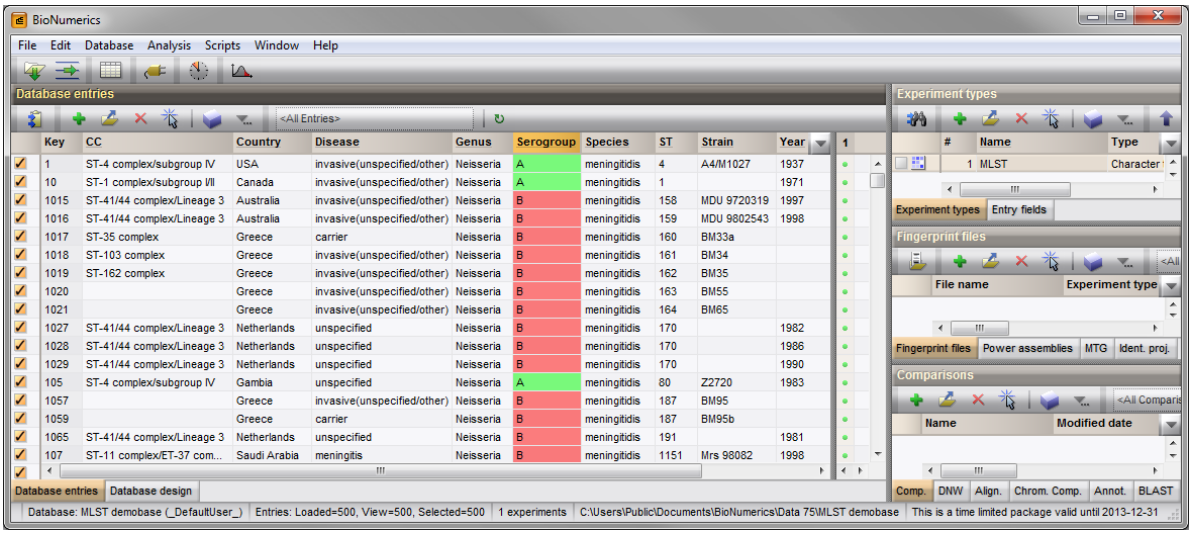

Figure 6: The *Main* window.

4. 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***.

5. In the *Group creation preferences* dialog box, make sure ***Create largest group 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 7).

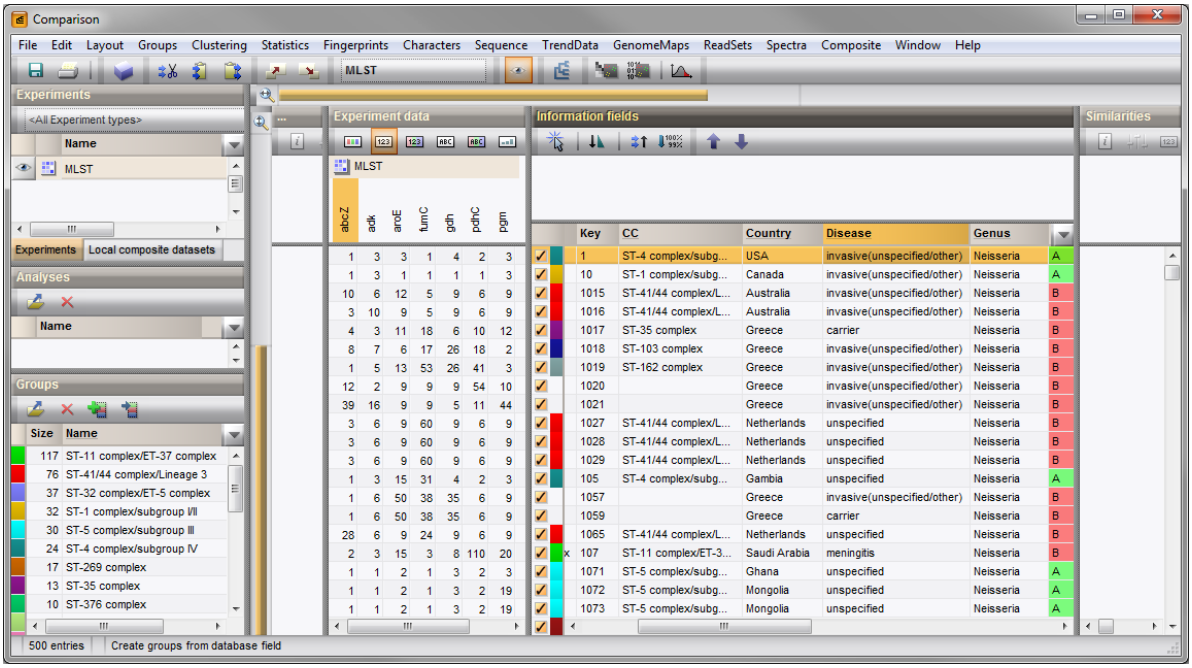

Figure 7: The *Comparison* window with groups defined.

5 Advanced clustering window

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

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

Due to the arbirariness 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. 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.

2. Specify an analysis name (for example **MLST1**), 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 (see 4) are automatically shown as node colors, but this can very easily be changed to a field state grouping defined in the *Main* window (see 2):

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

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

The node colors are updated according to the serogroups.

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

6. 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.

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

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

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

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

11. Press <***OK***> to apply the new settings.

The *Advanced cluster analysis* window should now look like Figure 8.

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.

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

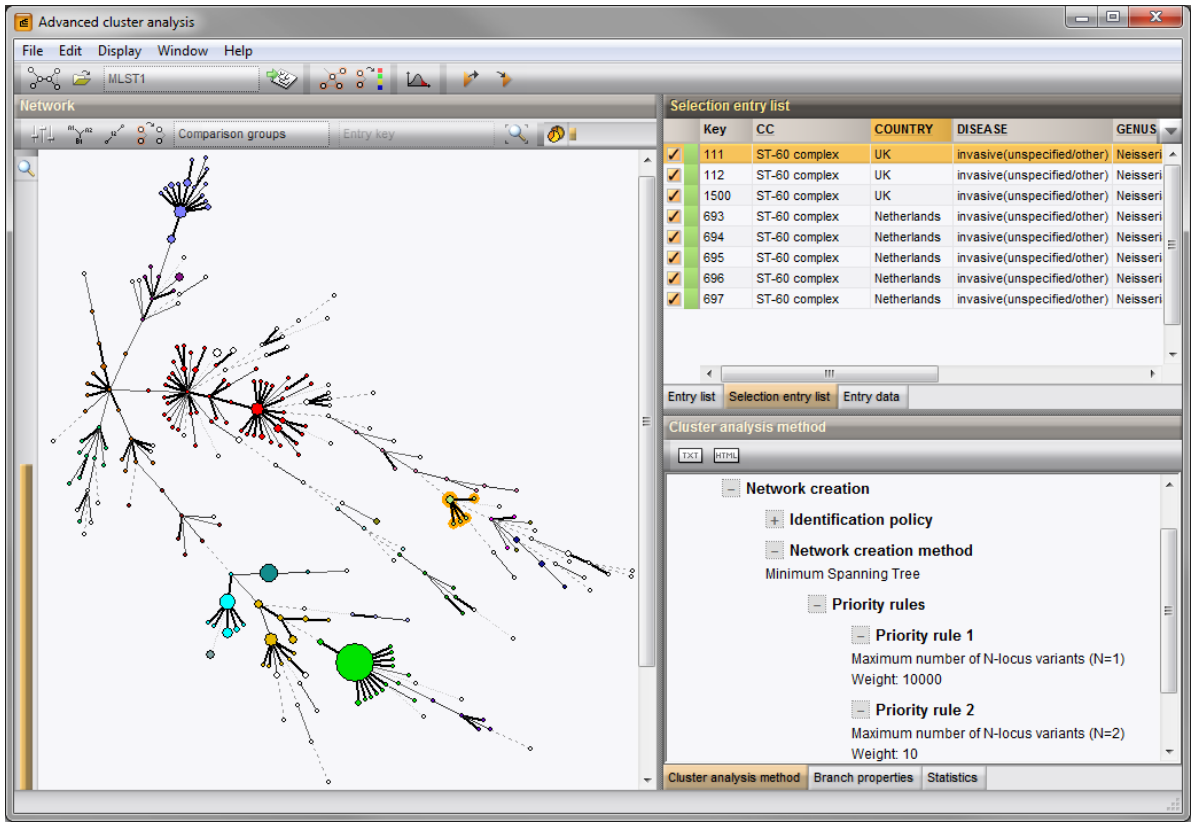

Figure 8: The *Advanced cluster analysis* window.

13. 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 result looks as in Figure 9. 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.

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.

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

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

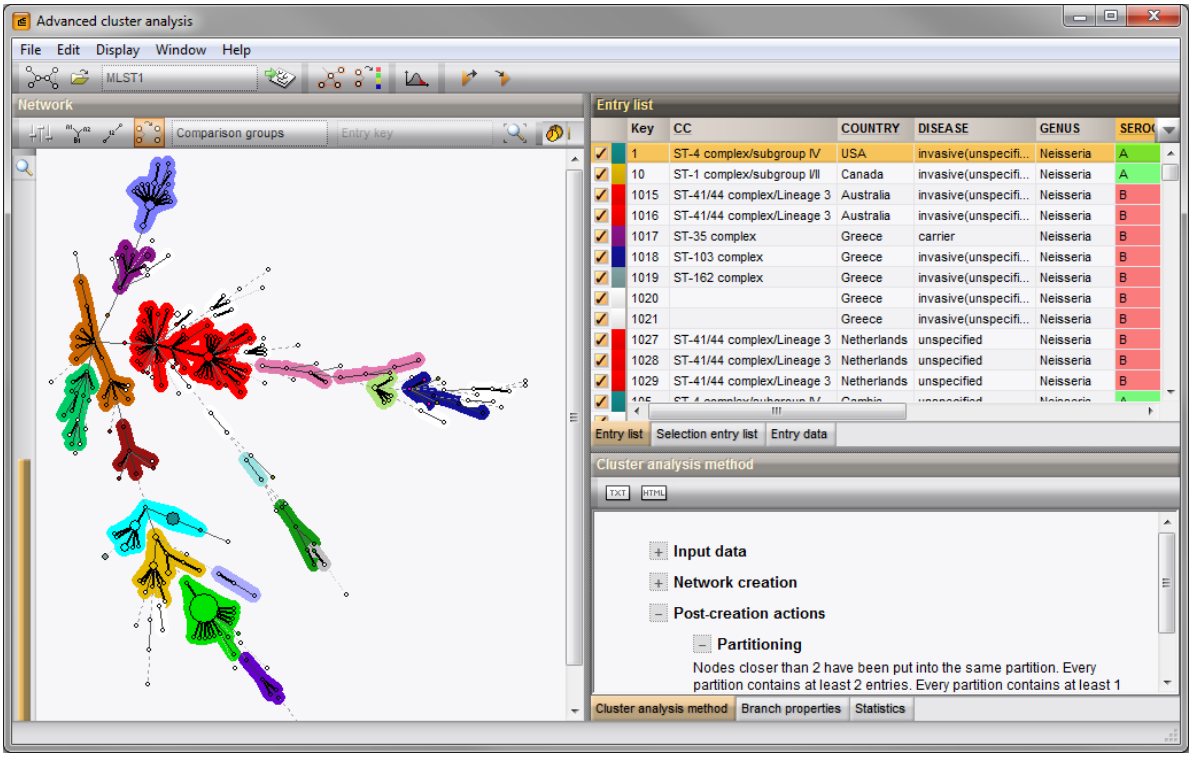

Figure 9: Partitions in the *Advanced cluster analysis* window.