BIONUMERICS Tutorial:

Cluster analysis of sequences

1 Aim

Similarity and distance-based trees (e.g. UPGMA and Neighbor joining) and phylogenetic trees (e.g. Maximum likelihood and Maximum parsimony trees) can be calculated in the *Comparison* window in BIONUMERICS, reflecting the relationships between the analyzed sequences. Different groups can be defined for the data and these groups are visualized in the trees, giving a nice overview of the relationships, the clusters present in the database and the outliers. In this tutorial we will calculate several types of trees on a set of ribosomal sequences.

2 Example data

As an example we will import sequences from the FASTA file `FastaSeqCL.txt` into a new or existing BIONUMERICS database. This FASTA file contains DNA sequences of the 16S rRNA gene for a set of 14 strains. The example file can be found on the download page on our website (`https://www.applied-maths.com/download/sample-data`, ”FASTA 16S sequences”).

3 Preparing the database

1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.

2. Select ***File*** > ***Import...*** (📂, **Ctrl+I**) to open the *Import* dialog box.

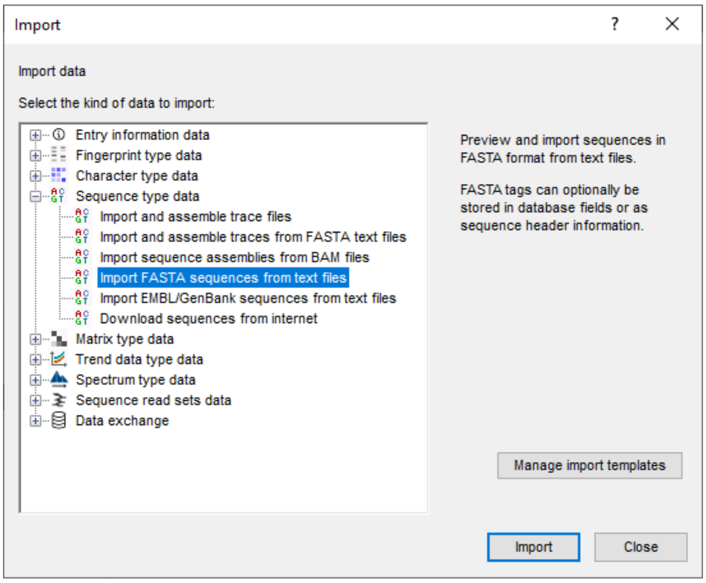
3. Choose the option ***Import FASTA sequences from text files*** under the ***Sequence type data*** item in the tree and click <***Import***> (see Figure 1).

4. The import wizard allows you to browse for one or more text files as data source. Press <***Browse***>, navigate to the folder, select the `FastaSeqCL.txt` file and press <***Open***>.

5. With the option ***Preview sequences*** checked, press <***Next***>.

The import wizard now displays a preview of the sequence data in the FASTA file (see Figure 2). From this preview, it is clear that the first FASTA field contains the unique strain number.

6. Press <***Next***>.



Import data

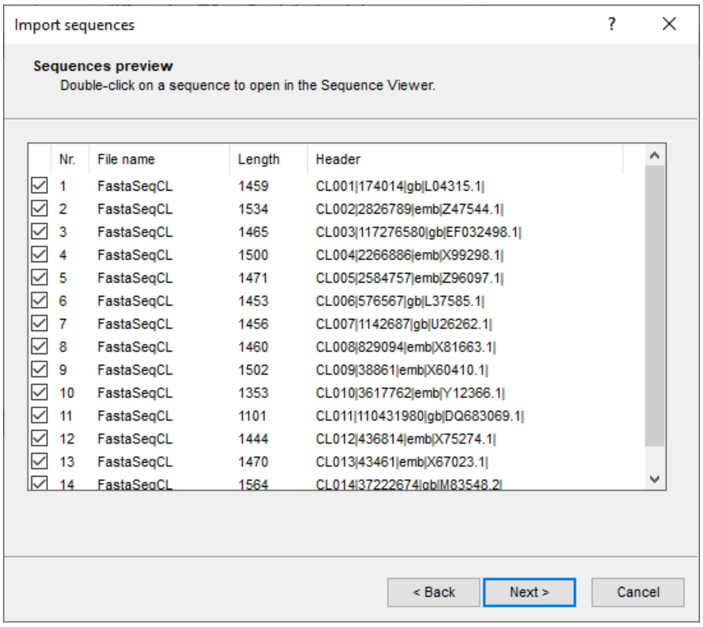
Select the kind of data to import:

- Entry information data
- Fingerprint type data
- Character type data
- Sequence type data
  - Import and assemble trace files
  - Import and assemble traces from FASTA text files
  - Import sequence assemblies from BAM files
  - Import FASTA sequences from text files
  - Import EMBL/GenBank sequences from text files
  - Download sequences from internet
- Matrix type data
- Trend data type data
- Spectrum type data
- Sequence read sets data
- Data exchange

Preview and import sequences in FASTA format from text files.

FASTA tags can optionally be stored in database fields or as sequence header information.

Figure 1: The Import tree.



Sequences preview

Double-click on a sequence to open in the Sequence Viewer.

	Nr.	File name	Length	Header
☑	1	FastaSeqCL	1459	CL001|174014|gb|L04315.1|
☑	2	FastaSeqCL	1534	CL002|2826789|emb|Z47544.1|
☑	3	FastaSeqCL	1465	CL003|117276580|gb|EF032498.1|
☑	4	FastaSeqCL	1500	CL004|2266886|emb|X99298.1|
☑	5	FastaSeqCL	1471	CL005|2584757|emb|Z96097.1|
☑	6	FastaSeqCL	1453	CL006|576567|gb|L37585.1|
☑	7	FastaSeqCL	1456	CL007|1142687|gb|U26262.1|
☑	8	FastaSeqCL	1460	CL008|829094|emb|X81663.1|
☑	9	FastaSeqCL	1502	CL009|38861|emb|X60410.1|
☑	10	FastaSeqCL	1353	CL010|3617762|emb|Y12366.1|
☑	11	FastaSeqCL	1101	CL011|110431980|gb|DQ683069.1|
☑	12	FastaSeqCL	1444	CL012|436814|emb|X75274.1|
☑	13	FastaSeqCL	1470	CL013|43461|emb|X67023.1|
☑	14	FastaSeqCL	1564	CL014|37222674|gb|M83548.2|

Figure 2: Sequence preview.

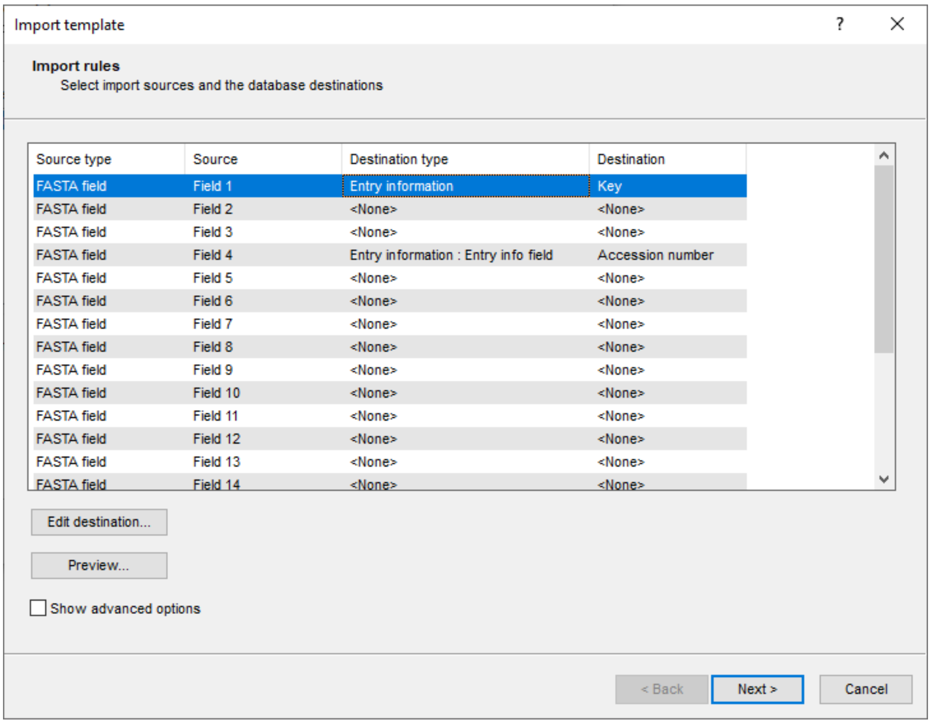
The next step of the import wizard lists the templates that are present to import sequence information in the database. As this is the first time we import FASTA formatted sequences in the database, we need to create a new import template by specifying ***Import rules***.

7. Click <***Create new***> to create a new import template.

8. Select "Field 1" in the list and click <***Edit destination***> or simply double-click on "Field 1". Select ***Key*** and press <***OK***>.

9. Double-click on "Field 4". Select ***Create new*** under ***Entry info field*** and click <***OK***>. Enter "Accession number" as name for the new information field, press <***OK***> and confirm the creation of the new field with <***Yes***>.

The grid is updated (see Figure 3).



Import template

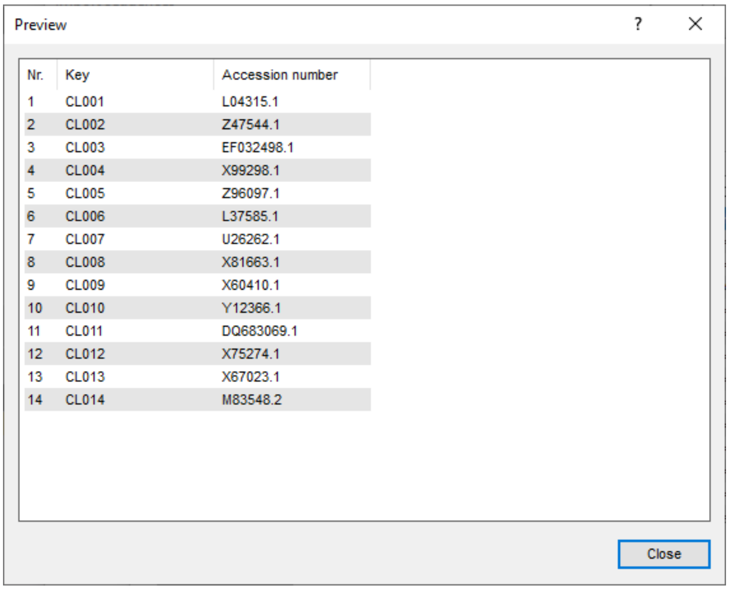
Import rules
Select import sources and the database destinations

Source type	Source	Destination type	Destination
FASTA field	Field 1	Entry information	Key
FASTA field	Field 2	<None>	<None>
FASTA field	Field 3	<None>	<None>
FASTA field	Field 4	Entry information : Entry info field	Accession number
FASTA field	Field 5	<None>	<None>
FASTA field	Field 6	<None>	<None>
FASTA field	Field 7	<None>	<None>
FASTA field	Field 8	<None>	<None>
FASTA field	Field 9	<None>	<None>
FASTA field	Field 10	<None>	<None>
FASTA field	Field 11	<None>	<None>
FASTA field	Field 12	<None>	<None>
FASTA field	Field 13	<None>	<None>
FASTA field	Field 14	<None>	<None>

□ Show advanced options

Figure 3: Import rules.

10. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import (see Figure 4).



Preview

Nr.	Key	Accession number
1	CL001	L04315.1
2	CL002	Z47544.1
3	CL003	EF032498.1
4	CL004	X99298.1
5	CL005	Z96097.1
6	CL006	L37585.1
7	CL007	U26262.1
8	CL008	X81663.1
9	CL009	X60410.1
10	CL010	Y12366.1
11	CL011	DQ683069.1
12	CL012	X75274.1
13	CL013	X67023.1
14	CL014	M83548.2

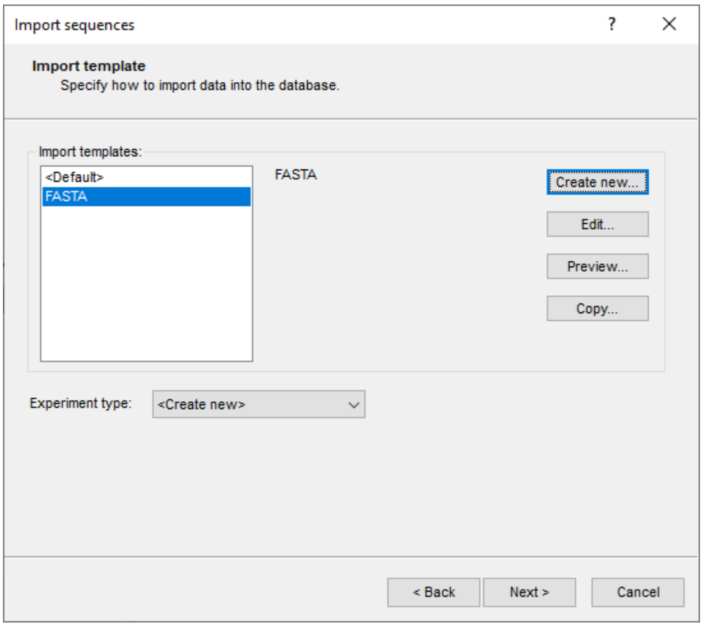
Figure 4: Preview of the rules.

11. Click <***Next***>.

12. Make sure ***Key*** is selected as ***Entry link field***. Press <***Finish***>.

13. Specify a template name, e.g. **FASTA**, and optionally enter a description. Press <***OK***>.

14. Highlight the newly created template and select "Create new" as ***Experiment type*** (see Figure 5).



Import sequences

Import template

Specify how to import data into the database.

Import templates:

- <Default>
- FASTA

FASTA

Experiment type: <Create new>

Figure 5: Import template.

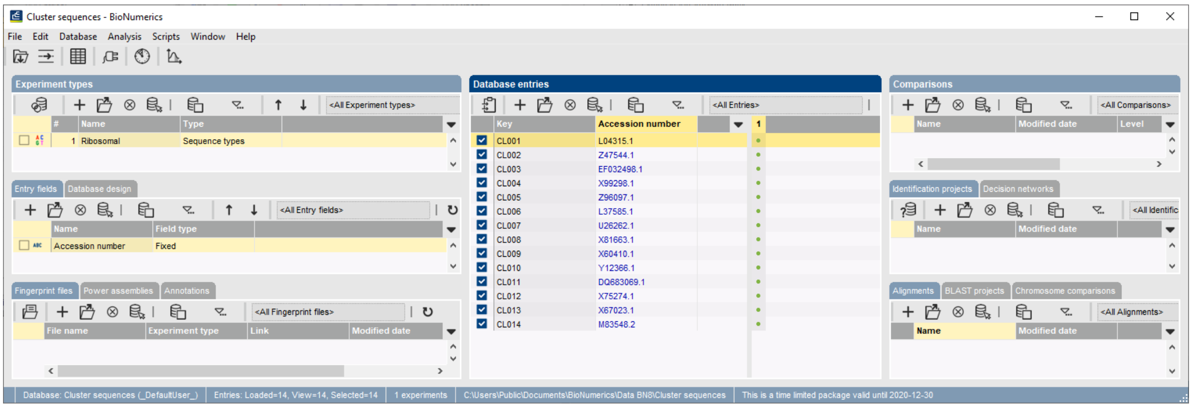
15. Press <***Next***>.

16. Specify a sequence type name (e.g. **Ribosomal**) and press <***OK***> and confirm the action.

The *Database links* wizard page will indicate that 14 new entries will be created during import.

17. Press <***Finish***> to start the import into the database.

For 14 strains, strain information and sequences are imported in the database (see Figure 6).



Cluster sequences - BioNumerics

Experiment types

#	Name	Type
1	Ribosomal	Sequence types

Entry fields

Name	Field type
Accession number	Fixed

Fingerprint files

File name	Experiment type	Link	Modified date

Database entries

	Key	Accession number		1
☑	CL001	L04315.1		●
☑	CL002	Z47544.1		●
☑	CL003	EF032498.1		●
☑	CL004	X99298.1		●
☑	CL005	Z96097.1		●
☑	CL006	L37585.1		●
☑	CL007	U26262.1		●
☑	CL008	X81663.1		●
☑	CL009	X60410.1		●
☑	CL010	Y12366.1		●
☑	CL011	DQ683069.1		●
☑	CL012	X75274.1		●
☑	CL013	X67023.1		●
☑	CL014	M83548.2		●

Comparisons

Name	Modified date	Level

Identification projects

Name	Modified date

Alignments

Name	Modified date

Database: Cluster sequences (_DefaultUser_) | Entries: Loaded=14, View=14, Selected=14 | 1 experiments | C:\Users\Public\Documents\BioNumerics\Data BN8\Cluster sequences | This is a time limited package valid until 2020-12-30

Figure 6: The *Main* window after import of the sequences.

4 Comparison window

1. In the *Database entries* panel of the *Main* window, select all entries that have an associated **Ribosomal** experiment. To select all entries at once, use the **CTRL+A** shortcut combination.

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 **Ribosomal** in the *Experiments* panel to display the sequences in the *Experiment data* panel.

Initially, the sequences are not aligned and no similarity matrix exists.

5 Pairwise and multiple alignment

1. The similarity matrix is calculated with ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** or by pressing and selecting ***Calculate cluster analysis***.

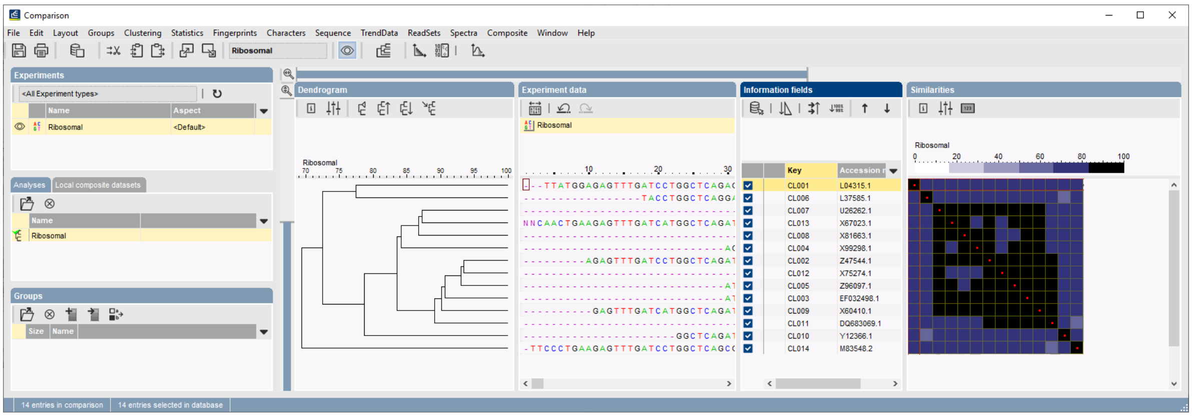
2. Choose ***Fast algorithm*** under ***Pairwise alignment*** with default settings and press <***Next***>.

3. In *Cluster analysis* wizard page, choose ***UPGMA*** as the dendrogram type and press <***Finish***>.

A dendrogram is now calculated, but the sequences are still unaligned.

4. Select ***Sequence*** > ***Multiple alignment...*** ( ).

5. Leave the settings as they are (default) and press <***OK***> to start the multiple alignment. When the calculations are done, the sequences are aligned in the *Experiment data* panel (see Figure 7).



		Key	Accession n
☑		CL001	L04315.1
☑		CL006	L37585.1
☑		CL007	U26262.1
☑		CL013	X67023.1
☑		CL008	X81663.1
☑		CL004	X99298.1
☑		CL002	Z47544.1
☑		CL012	X75274.1
☑		CL005	Z96097.1
☑		CL003	EF032498.1
☑		CL009	X60410.1
☑		CL011	DQ683069.1
☑		CL010	Y12366.1
☑		CL014	M83548.2

Figure 7: Multiple alignment.

In order to facilitate visual interpretation of global alignments there are three methods to highlight homologous regions. The first one is the *Neighbor match* representation:

6. Select ***Sequence*** > ***Block type*** > ***Neighbor blocks*** to show the Neighbor match representation (see Figure 8).

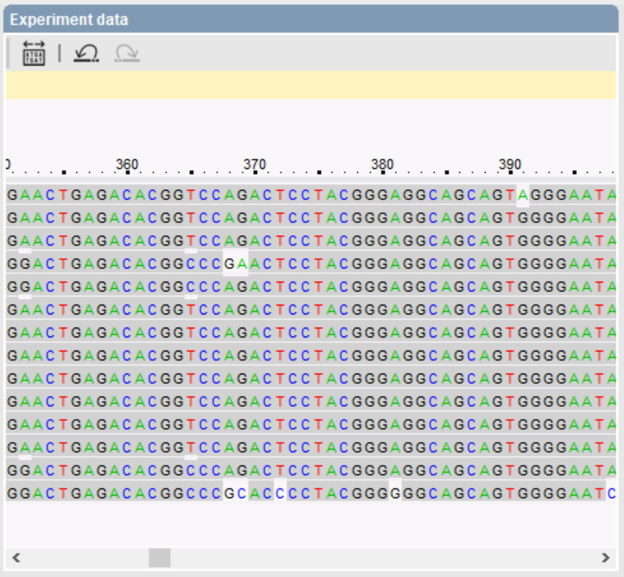
This representation shows bases as blocks (highlighted) if at least one of the neighboring sequences has the same base at the corresponding position. Between two different groups of consensus, a small line is drawn.

The second visualization, the *Consensus match*, first requires a consensus sequence to be present.

7. Select the root and select ***Sequence*** > ***Create consensus of branch***.

8. Leave the default setting and press <***OK***>.

A consensus sequence of the root is shown on the header of the image panel. Bases for which there is a consensus in more than 50% of the sequences, are named in the consensus, the other

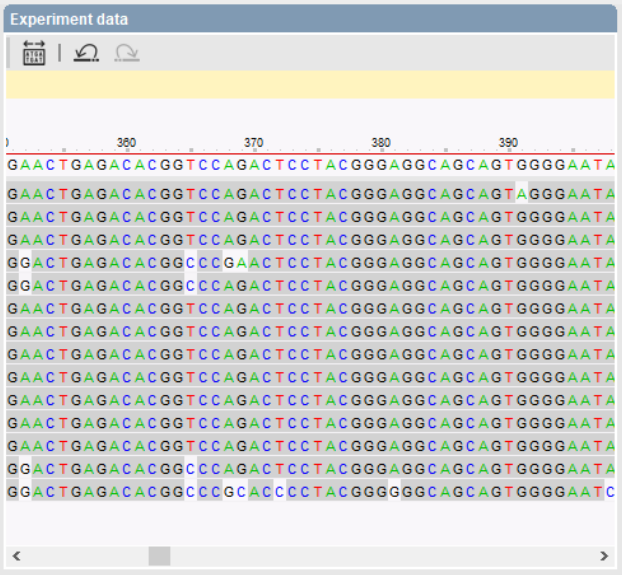


```
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCGAACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCGCACCCCTACGGGGGGCAGCAGTGGGGAATC
```

Figure 8: Neighbor match representation.

bases are unnamed (N).

9. Select ***Sequence*** > ***Block type*** > ***Consensus blocks*** to show the consensus match representation (see Figure 9).



```
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA

GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCGAACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATA
GGACTGAGACACGGCCCGCACCCCTACGGGGGGCAGCAGTGGGGAATC
```

Figure 9: Consensus match representation.

A third method, the *Consensus difference*, displays the consensus sequence in the editor caption, and only shows bases that differ from the consensus. Bases that are the same as the consensus are shown as "|" (see Figure 10).

10. Select ***Sequence*** > ***Block type*** > ***Consensus difference***.

A multiple alignment can be edited manually and is saved along with the comparison.

11. Select ***File*** > ***Save*** (🖫, **Ctrl+S**) to save the multiple alignment.

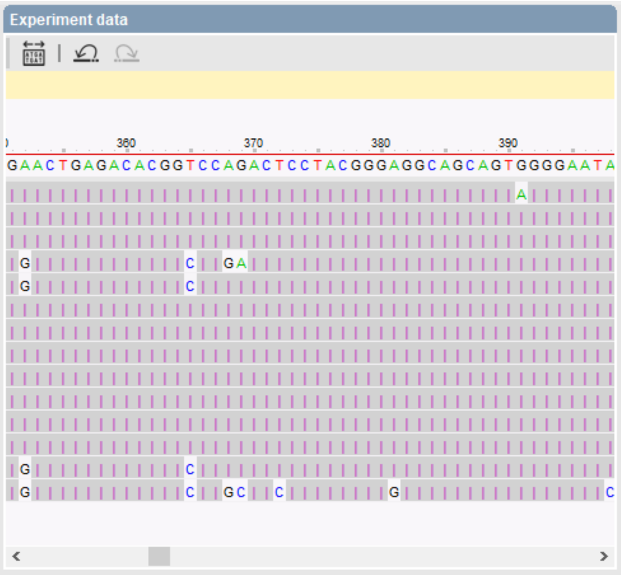


Figure 10: Consensus difference representation.

12. A window pops up asking you to enter a name for the comparison. Enter “SeqAlign” and press <*OK*>.

6 Cluster analysis

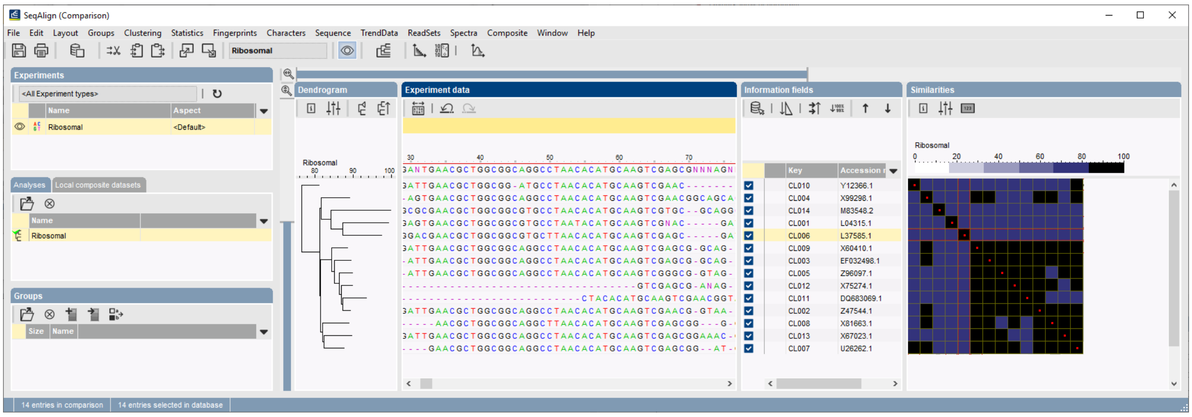
6.1 Similarity or distance-based clustering

The mutual similarities between all the sequences are calculated from the aligned sequences as present in the multiple alignment.

1. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** or press and select ***Calculate cluster analysis***.

2. Choose ***Multiple alignment based*** with the default settings and press <***Next***>.

3. In the second step of the wizard, select ***Neighbor Joining*** as dendrogram type and click <***Finish***>.



	Key	Accession n
☑	CL010	Y12366.1
☑	CL004	X99298.1
☑	CL014	M83548.2
☑	CL001	L04315.1
☑	CL006	L37585.1
☑	CL009	X60410.1
☑	CL003	EF032498.1
☑	CL005	Z96097.1
☑	CL012	X75274.1
☑	CL011	DQ683069.1
☑	CL002	Z47544.1
☑	CL008	X81663.1
☑	CL013	X67023.1
☑	CL007	U26262.1

Figure 11: Neighbor joining based on multiple alignment.

4. Save the comparison with ***File*** > ***Save*** (, **Ctrl+S**).

6.2 Phylogenetic clustering

6.2.1 Introduction

The most widely used phylogenetic clustering methods are maximum parsimony (Fitch, 1971) and maximum likelihood (Felsenstein, 1981). In both methods, an evolution is reconstructed based upon sequence data by optimizing a certain criterion.

The *maximum parsimony* method tries to find a tree that explains the sequence diversity with a minimum number of total mutations needed (i.e. the most parsimonious tree). The branch lengths of the tree reflect the number of mutations along the branches.

The *maximum likelihood* method is based on a probabilistic model for base substitution. A tree is searched for that has the highest likelihood, i.e. the probability that the given sequences are the result of an evolution along that tree, following the assumed probabilistic module. The branch lengths of the tree correspond to evolutionary time.

The Neighbor joining method can also be classified under phylogenetic clustering methods, but in this tutorial it is covered in 6.1 since it uses a distance matrix as input.

6.2.2 Maximum parsimony

5. In the *Comparison* window with the multiple alignment calculated press the <***F4***> to clear the selection.

6. Select CL007, CL008, CL013 using the check boxes and choose ***Groups*** > ***Create new group from selection*** (, **Ctrl+G**). Specify a group name (e.g. **Group 1**) and press <***OK***> to confirm the creation of the group.

7. Press <***F4***> again to clear the selection.

8. Select CL002, CL003, CL005, CL009, CL011, CL012, and choose ***Groups*** > ***Create new group from selection*** (, **Ctrl+G**). Specify a group name (e.g. **Group 2**) and press <***OK***> to confirm the creation of the group.

9. Press <***F4***> again to clear the selection.

10. Select CL010 and add the entry to a third group (e.g. **Group 3**).

11. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...***.

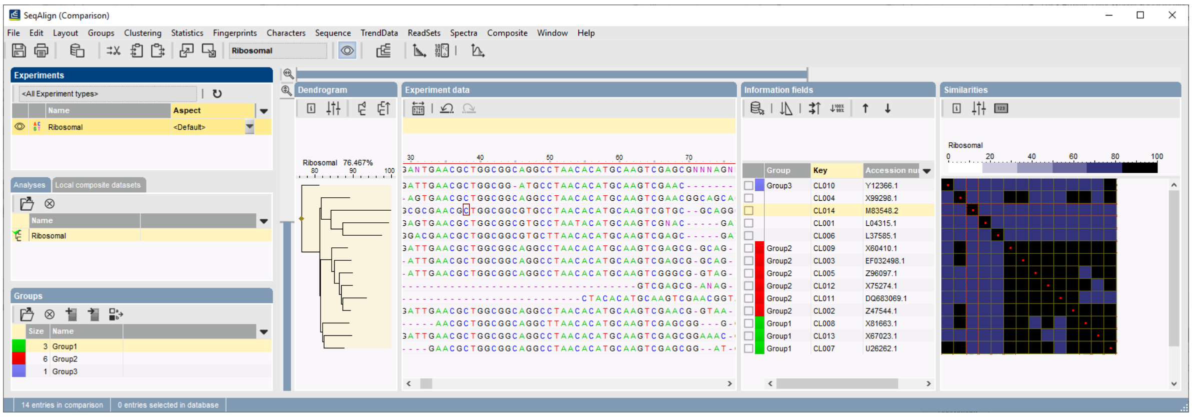
12. Name the cluster analysis **Max parsimony**, choose ***Ribosomal*** as experiment and ***Maximum parsimony tree*** as predefined template (see Figure 13).

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

13. Press <***Next***> to start the calculations.

The resulting tree is displayed in *Advanced cluster analysis* window (see Figure 14). The parsimony (the total number of base conversions over the tree) is given in the *Statistics* tab. The entries are represented in the colors of the groups we have defined earlier.

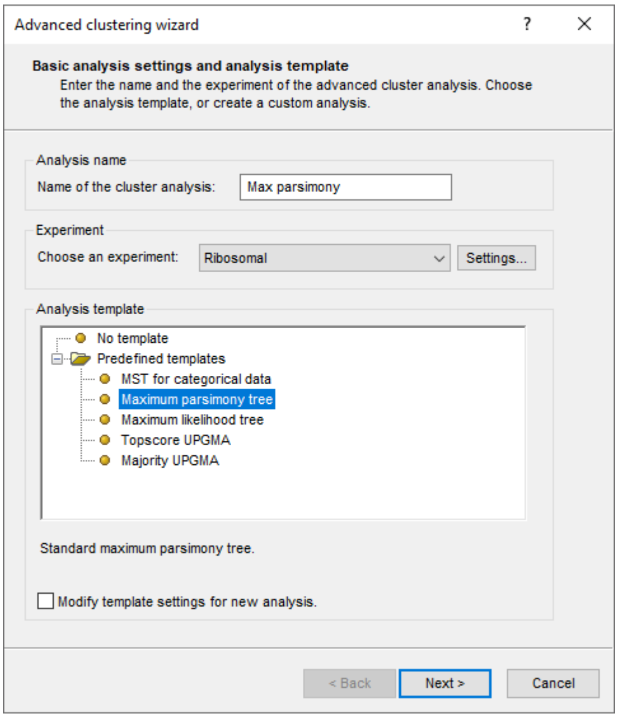
14. Select ***Display*** > ***Display settings***, check ***Show node labels*** in the *Node labels and sizes* tab, and check ***Show branch labels*** in the *Branch labels and sizes* tab. Press <***OK***>.



Size	Name
🟩 3	Group1
🟥 6	Group2
🟦 1	Group3

	Group	Key	Accession nu
□ 🟦	Group3	CL010	Y12366.1
□		CL004	X99298.1
□		CL014	M83548.2
□		CL001	L04315.1
□		CL006	L37585.1
□ 🟥	Group2	CL009	X60410.1
□ 🟥	Group2	CL003	EF032498.1
□ 🟥	Group2	CL005	Z96097.1
□ 🟥	Group2	CL012	X75274.1
□ 🟥	Group2	CL011	DQ683069.1
□ 🟥	Group2	CL002	Z47544.1
□ 🟩	Group1	CL008	X81663.1
□ 🟩	Group1	CL013	X67023.1
□ 🟩	Group1	CL007	U26262.1

Figure 12: The *Comparison* window with groups defined.



Basic analysis settings and analysis template

Enter the name and the experiment of the advanced cluster analysis. Choose the analysis template, or create a custom analysis.

Analysis name

Name of the cluster analysis: Max parsimony

Experiment

Choose an experiment: Ribosomal

Settings...

Analysis template

- No template
- Predefined templates
  - MST for categorical data
  - Maximum parsimony tree
  - Maximum likelihood tree
  - Topscore UPGMA
  - Majority UPGMA

Standard maximum parsimony tree.

□ Modify template settings for new analysis.

< Back | Next > | Cancel

Figure 13: Maximum parsimony template.

15. Select ***Display*** > ***Zoom to fit*** to optimize the view of the tree in the current window.

16. By adjusting the zoom bar, the node sizes can be increased/decreased.

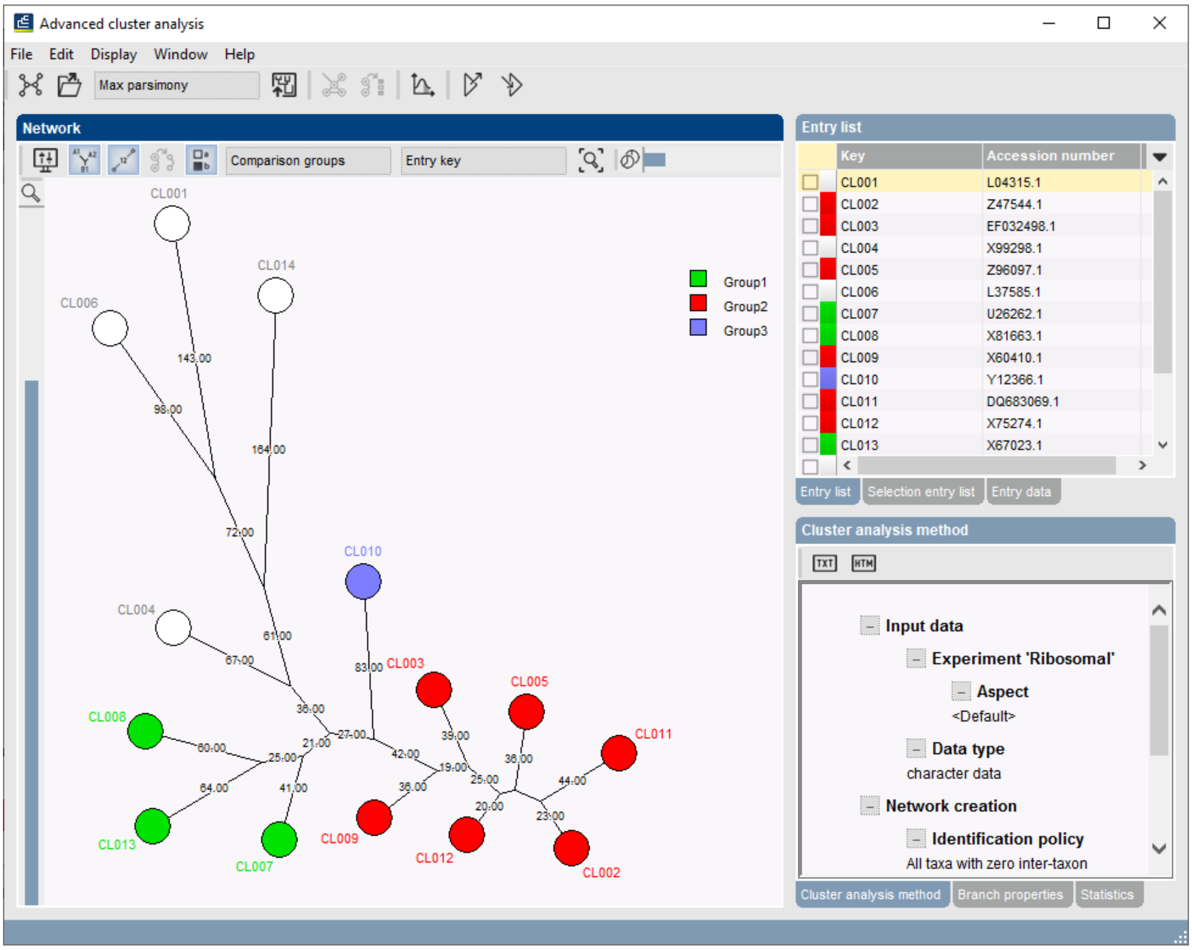
The right bottom information panel displays the cluster analysis settings, the branch properties and the total parsimony of the tree in the statistics tab.

17. A node or branch can be selected by clicking on them, or several nodes/branches by holding the **Shift**-key while clicking.

We will now perform a bootstrap analysis on the Maximum Parsimony tree.

18. Go to ***Edit*** > ***Compute statistics*** to call the *Statistics* dialog.

19. Check ***Perform resampling***, set the resampling strategy on ***Bootstrap resampling*** with 100 samples and click <***OK***>.



Entry list

	Key	Accession number
□	CL001	L04315.1
□ 🟥	CL002	Z47544.1
□ 🟥	CL003	EF032498.1
□	CL004	X99298.1
□ 🟥	CL005	Z96097.1
□	CL006	L37585.1
□ 🟩	CL007	U26262.1
□ 🟩	CL008	X81663.1
□ 🟥	CL009	X60410.1
□ 🟦	CL010	Y12366.1
□ 🟥	CL011	DQ683069.1
□ 🟥	CL012	X75274.1
□ 🟩	CL013	X67023.1

Cluster analysis method

- Input data
  - Experiment 'Ribosomal'
    - Aspect
      <Default>
  - Data type
    character data
- Network creation
  - Identification policy
    All taxa with zero inter-taxon

Figure 14: Maximum parsimony tree.

We will now display a rooted tree from the unrooted tree.

20. Select the CL010 entry to act as root, go to ***Edit*** > ***Determine root*** and choose ***Use selected node as root position***.

Once a rooted tree has been calculated, it can be visualized in the *Comparison* window:

21. Select ***File*** > ***Show dendrogram in comparison*** and confirm.

In a phylogenetic context, a dendrogram is sometimes represented as a real tree with a stem and branches:

22. This can be achieved by selecting ***Clustering*** > ***Display rendered tree*** in the *Comparison* window and pressing <***OK***> (see Figure 15).

23. Close the rendered tree representation.

6.2.3 Maximum likelihood

To reduce the calculation time of the maximum likelihood tree we will delete the 4 entries that do not belong to a group:

24. Press the F4 key in the *Comparison* window to clear the selection. Select entry CL004, CL006, CL001 and CL014 using the check boxes and select ***Edit*** > ***Cut selection*** (✂, **Ctrl+X**). Confirm the deletion of the four entries from the comparison.

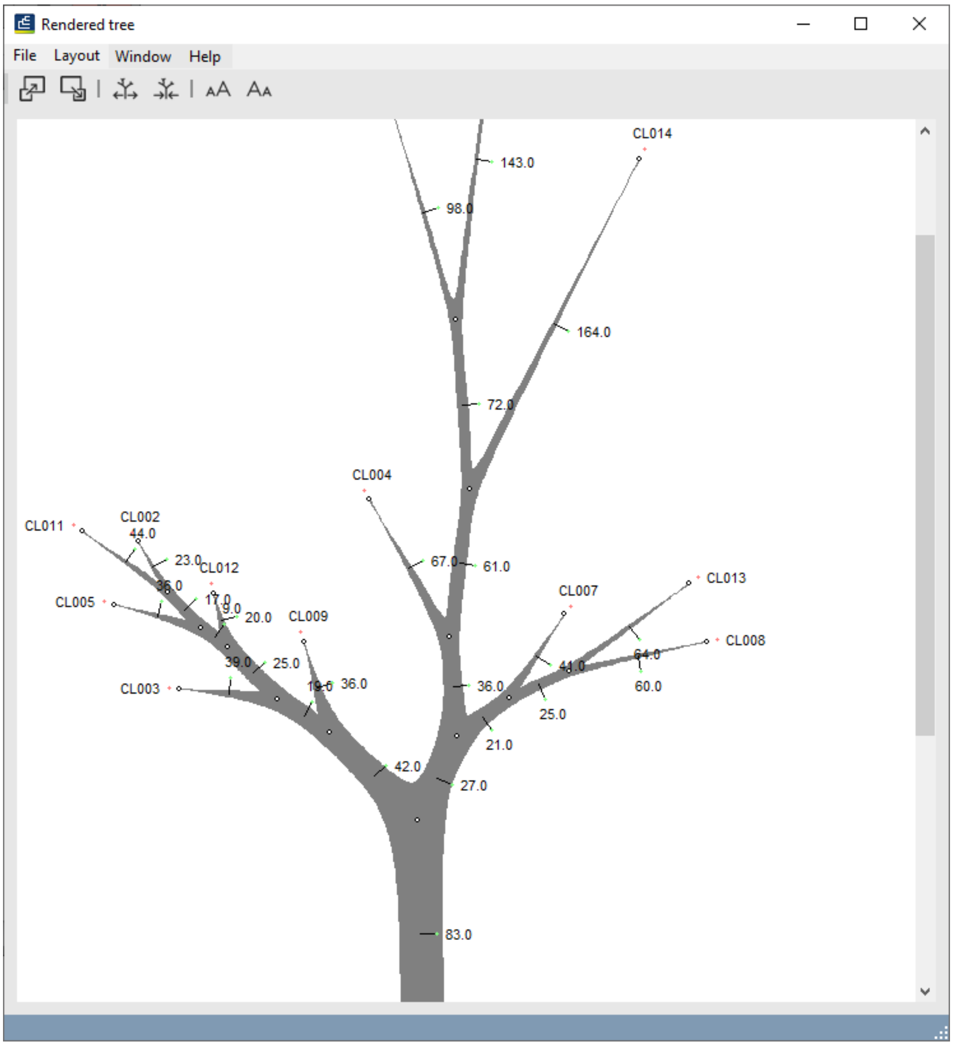


Figure 15: Rendered tree representation.

25. Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...***.

26. Name the cluster analysis **Max likelihood**, choose ***Ribosomal*** as experiment and ***Maximum likelihood tree*** as predefined template (see Figure 16).

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

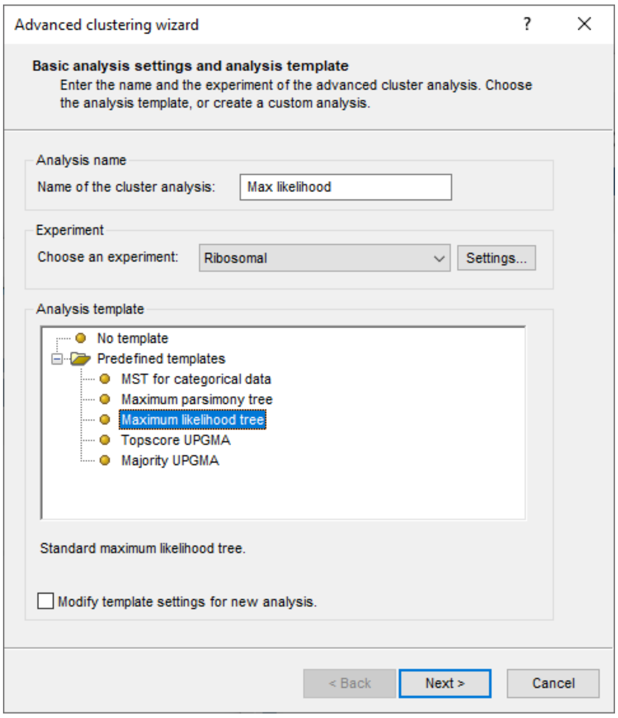
27. Press <***Next***> to start the calculations.

The resulting tree is displayed in *Advanced cluster analysis* window (see Figure 17). The entries are represented in the colors of the groups we have defined earlier.

28. Select ***Display*** > ***Display settings***, check ***Show node labels*** in the *Node labels and sizes* tab. Press <***OK***>.

29. Select ***Display*** > ***Zoom to fit*** to optimize the view of the tree in the current window.

30. By adjusting the zoom bar, the node sizes can be increased/decreased.



Basic analysis settings and analysis template

Enter the name and the experiment of the advanced cluster analysis. Choose the analysis template, or create a custom analysis.

Analysis name

Name of the cluster analysis: Max likelihood

Experiment

Choose an experiment: Ribosomal

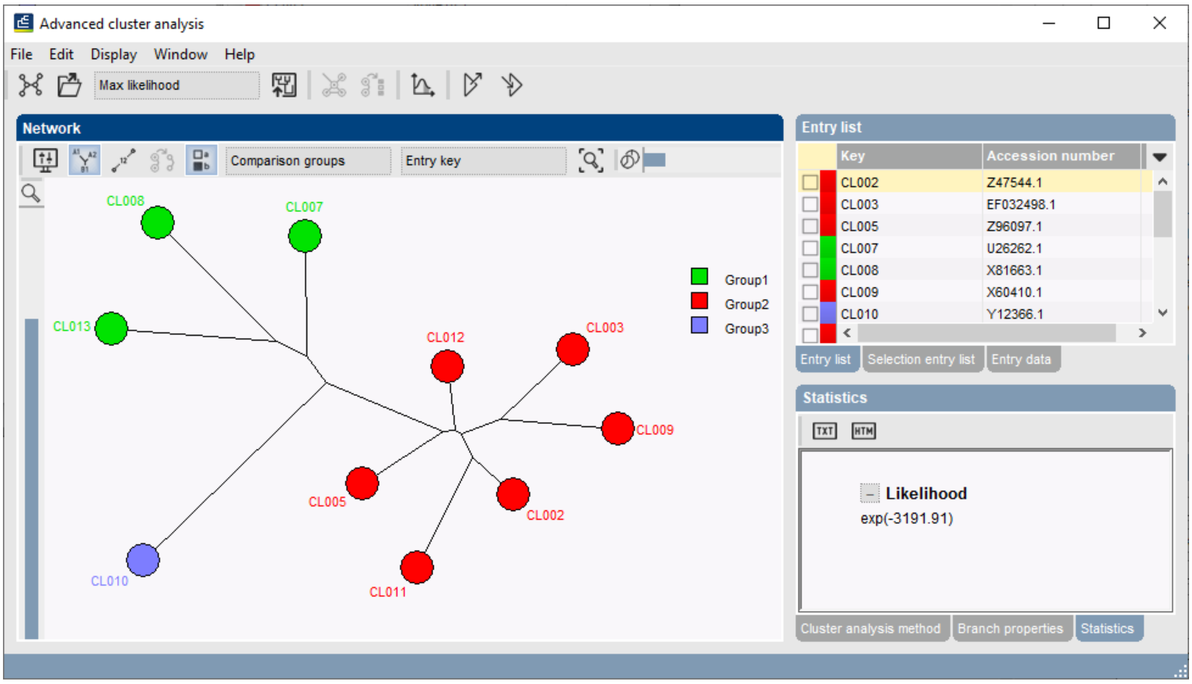
Analysis template

- ● No template
- Predefined templates
  - ● MST for categorical data
  - ● Maximum parsimony tree
  - ● Maximum likelihood tree
  - ● Topscore UPGMA
  - ● Majority UPGMA

Standard maximum likelihood tree.

☐ Modify template settings for new analysis.

Figure 16: Maximum likelihood template.



		Key	Accession number
☐	🟥	CL002	Z47544.1
☐	🟥	CL003	EF032498.1
☐	🟥	CL005	Z96097.1
☐	🟩	CL007	U26262.1
☐	🟩	CL008	X81663.1
☐	🟥	CL009	X60410.1
☐	🟦	CL010	Y12366.1

Likelihood

exp(-3191.91)

Figure 17: Maximum likelihood tree.