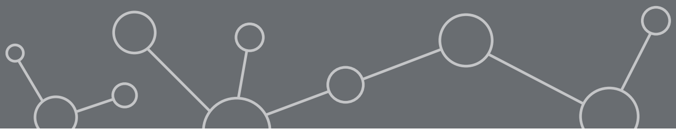
BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

MLST online plugin

PLUGINS

VERSION 7.6



www.applied-maths.com

Contents

NOTES

SUPPORT BY APPLIED MATHS

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of BioNumerics®, or suggestions for improvement, refinement or extension of the software to your specific applications:

Applied Maths NV
Keistraat 120
9830 Sint-Martens-Latem
Belgium
PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: info@applied-maths.com
URL: http://www.applied-maths.com

Applied Maths, Inc.
11940 Jollyville Road, Suite 115N
Austin, Texas 78759
U.S.A.
PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: info-US@applied-maths.com

LIMITATIONS ON USE

The BioNumerics® software, its plugin tools and their accompanying guides are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement. No part of this guide may be reproduced by any means without prior written permission of the authors.

BioNumerics® uses following third-party software tools and libraries:

- The Python® 2.7.4 release from the Python Software Foundation (http://www.python.org/).
- A library for XML input and output from the Apache Software Foundation (http://www.apache.org).
- NCBI toolkit version 2.2.10 (http://www.ncbi.nlm.nih.gov/BLAST/).
- The Boost c++ libraries (http://www.boost.org/).
- Samtools for interacting with SAM / BAM files (http://www.htslib.org/download/)
- The 7-Zip command line version (7za.exe) from 7-Zip, copyright 1999-2010 Igor Pavlov. http://www.7-zip.org/
- Velvet for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Ray for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Mothur for Windows, source code can be downloaded from http://www.applied-maths.com/download/open-source.
- Cairo 2D graphics library version 1.12.14 (http://cairographics.org/).
- Crypto++ Library version 5.5.2 (http://www.cryptopp.com/).
- libSVM library for Support Vector Machines (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).
- SQLite version 3.7.17 (http://www.sqlite.org/).
- Gecko engine version 21 (https://developer.mozilla.org/en-US/docs/Mozilla/Gecko).
- pymzML Python® module for high throughput bioinformatics on mass spectrometry data (https://github.com/pymzml/pymzML).
- Numpy Python® library version 1.8.1 (http://www.numpy.org/).
- BioPython Python® library version 1.64 (http://www.biopython.org/).
- PIL Python library® version 1.1.7 (http://www.pythonware.com/products/pil/).
- The SPAdes genome assembler version 3.7.1 (http://bioinf.spbau.ru/spades).

Chapter 1

Starting and setting up BioNumerics

1.1 Introduction

This guide is designed as a tutorial for the *Multi Locus Sequence Typing plugin* of BioNumerics. With the *MLST online plugin* you can:

1. Download the allele numbering, sequence types and clonal complex information for a selected organism from a large number of public online repositories (e.g. PubMLST.org, MLST.net, . . .) and query this information.
2. Query the databases available on the PubMLST.org (`http://pubmlst.org`) directly without using the standard web interface.
3. Store custom allele numbering, sequence type and clonal complex information in the BioNumerics database, and query this information.

The minimal configuration for the installation of the *MLST online plugin* includes the Sequence data module (import and storage of sequences) and the Character data module (storage of allelic profiles).

1.2 Startup program

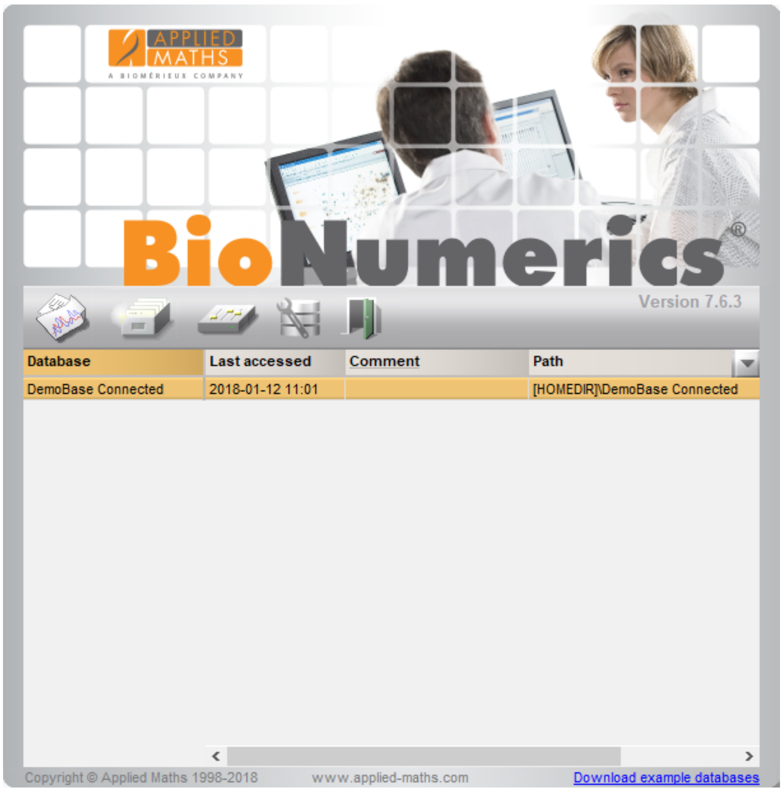
When BioNumerics is launched from the Windows start panel or when the BioNumerics shortcut () on your computer's desktop is double-clicked, the **Startup program** is run. This program shows the *BioNumerics Startup* window (see Figure 1.1).

A new BioNumerics database is created from the Startup program by pressing the button.

An existing database is opened in BioNumerics with or by simply double-clicking on a database name in the list.

1.3 Creating a new database

3.1 Press the button in the BioNumerics *BioNumerics Startup* window to enter the *New database* wizard.



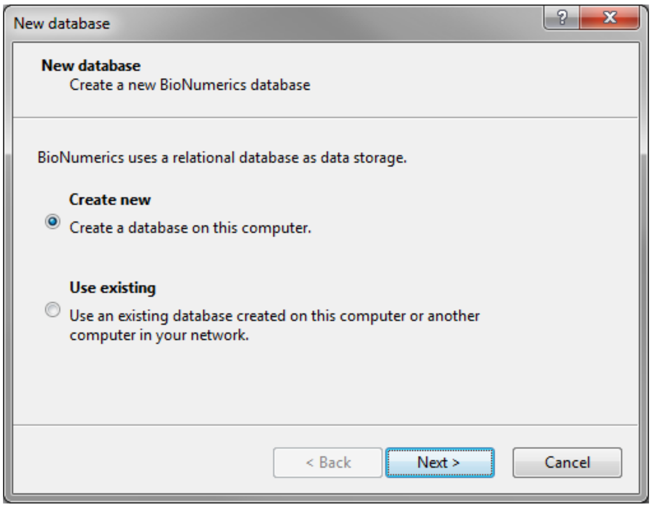
Database	Last accessed	Comment	Path
DemoBase Connected	2018-01-12 11:01		[HOMEDIR]\DemoBase Connected

Figure 1.1: The *BioNumerics Startup* window.

3.2 Enter a name for the database, and press <***Next***>.

A new dialog box pops up, prompting for the type of database (see Figure 1.2).

3.3 Since we want to create a new database to demonstrate the features of the plugin, leave the default option selected and press <***Next***>.



New database
Create a new BioNumerics database

BioNumerics uses a relational database as data storage.

Create new
● Create a database on this computer.

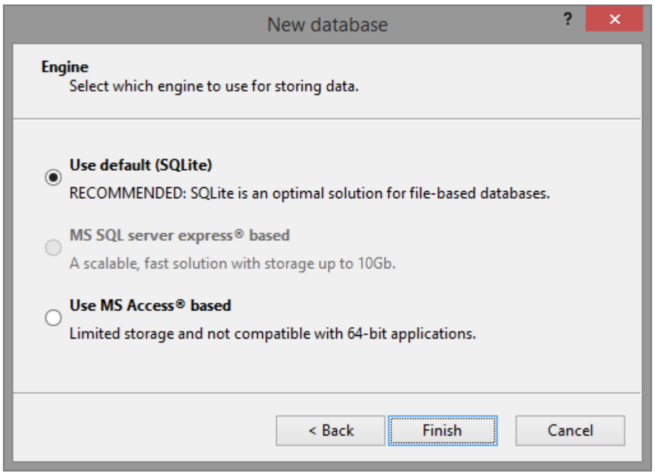
Use existing
○ Use an existing database created on this computer or another computer in your network.

Figure 1.2: The *New database* wizard page.

A new dialog box pops up, prompting for the database engine (see Figure 1.3).

3.4 Leave the default option selected and press <***Next***>.

3.5 Press <***Finish***> to complete the setup of the new database.



Engine
Select which engine to use for storing data.

● **Use default (SQLite)**
RECOMMENDED: SQLite is an optimal solution for file-based databases.

○ **MS SQL server express® based**
A scalable, fast solution with storage up to 10Gb.

○ **Use MS Access® based**
Limited storage and not compatible with 64-bit applications.

Figure 1.3: The *Database engine* wizard page.

The *Plugins* dialog box appears.

Chapter 2

Installing the MLST online plugin

2.1 MLST setup

Before an organism can be typed with MLST, an *MLST scheme* needs to be defined for that organism. An MLST scheme consists of a list of loci (typically seven), which are sequenced and assigned an allele number. Furthermore, in order to achieve an unambiguous nomenclature, a list of *allele variants* per locus needs to be provided and a list of *allelic profiles* corresponding to *sequence types*.

The *MLST online plugin* can be set up in different ways, depending on the available information for the organism of interest. An MLST scheme is defined by its **profile and allele files**: a tab-delimited text file with the sequence types and their corresponding allelic profiles and, for each of the loci, a FASTA-formatted text file with the allele variant definitions. These files can be made available on the local computer, the network or on the internet. A path or URL should be specified for each of the files. Even if no MLST scheme is publicly available, the plugin can be used to type organisms using MLST.

When a MLST data repository is available online for the organism of your interest, the MLST setup comes down to simply picking the organism from the provided list. The profile and allele variant definitions will be downloaded and stored locally in the BioNumerics database. This information can be automatically updated from the online definition files.

It can occur that your organism is not listed (meaning that it is not available from a known online data repository), but nevertheless an MLST scheme has been published for this organism and profile and allele files are available. In this case, the organism name, locus names and the locations of the profile and allele files (a path on the local computer, network drive or URL) should be entered manually. The profile and allele definitions will then be downloaded and can be automatically updated from the definition files.

In case no MLST scheme has been published yet for the organism of your interest, the only option is to store all data locally and to update the profile and allele definitions manually.

Once the *MLST online plugin* is installed, it will not be possible anymore to modify the MLST scheme (organism name, locus names) for that database. It is possible, however, to change the way the MLST data is updated. For example, the following scenario can be accommodated for:

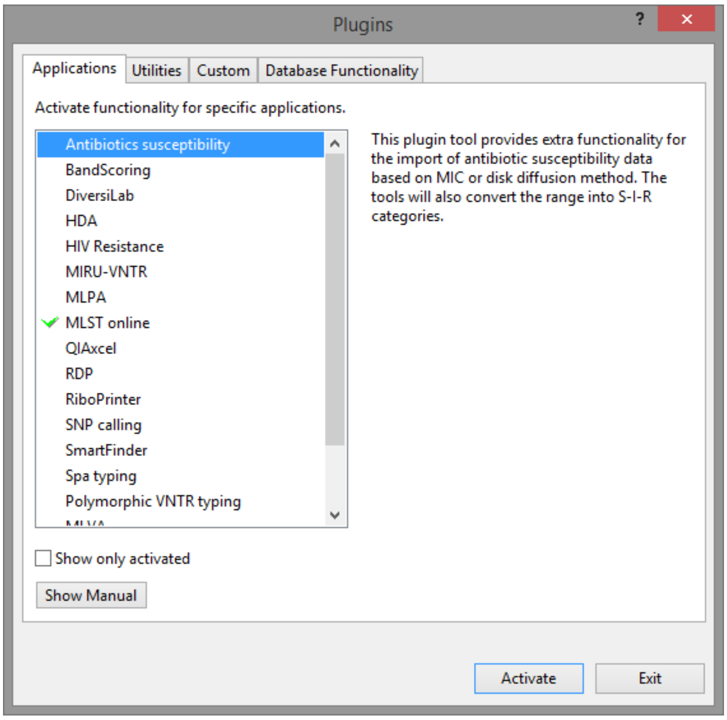
1. Startup of a new MLST scheme: all data is stored locally; manual access only
2. Share the MLST scheme between different databases via local files (tools are available to export the profile and allele definitions to text files)
3. Make the MLST scheme public by putting the definition files on the web

The following paragraphs will guide you through the *MLST online plugin* installation process.

2.2 The Plugin installation toolbox

If a database is opened for the first time, the *Plugins* dialog box will appear by default (see Figure 2.1).

If the database has already been opened previously, the *Plugins* dialog box can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** (🔌).



Activate functionality for specific applications.

This plugin tool provides extra functionality for the import of antibiotic susceptibility data based on MIC or disk diffusion method. The tools will also convert the range into S-I-R categories.

☐ Show only activated

Figure 2.1: The *Plugins* dialog box.

When a particular plugin is selected from the list of plugins, a short description appears in the right panel.

A selected plugin can be installed with the <***Activate***> button. The software will ask for confirmation before installation. Some plugins depend on functionality offered by specific BioNumerics modules. If a required module is missing, the plugin cannot be installed and an error message will be generated.

Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the <***Deactivate***> button.

If the selected plugin is documented, pressing <***Show Manual***> will open its manual in the *Help* window.

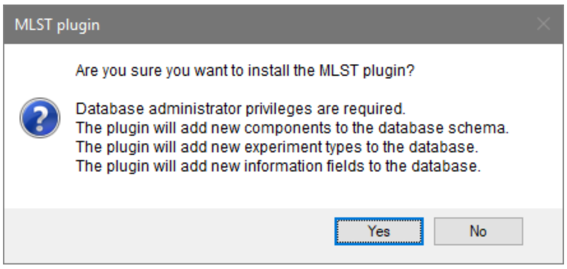
2.1 Select the *MLST online plugin* from the list in the *Applications tab* and press the <***Activate***> button.

BioNumerics asks the user to confirm the installation of the *MLST online plugin* (see Figure 2.2). Installation of the plugin requires administrator privileges on the relational database.

2.2 Press <***Yes***> to start with the installation.

In the first step of the wizard, the user is asked to select the organism source (see Figure 2.3).

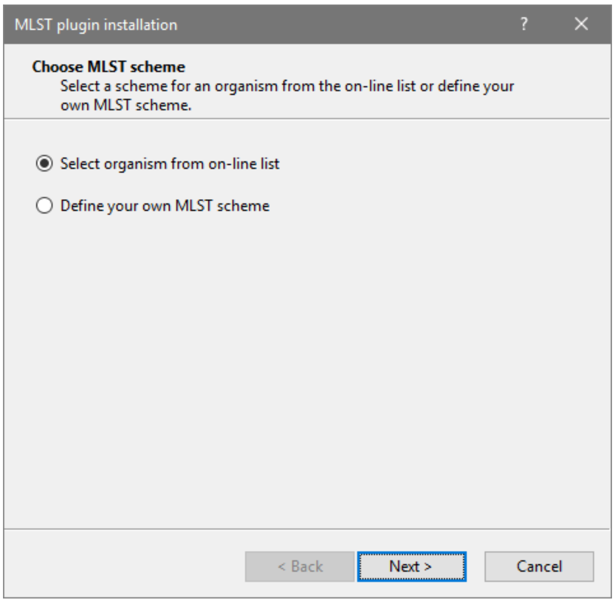
- ***Select organism from on-line list***: Allele numbering, sequence type and clonal complex information from online repositories is used for typing. All available organisms are listed in the next step of the wizard (see 2.3).
- ***Define your own MLST scheme***: Custom information is stored in the BioNumerics database (see 2.4).



Are you sure you want to install the MLST plugin?

Database administrator privileges are required.
The plugin will add new components to the database schema.
The plugin will add new experiment types to the database.
The plugin will add new information fields to the database.

Figure 2.2: Confirmation message that appears when installing the *MLST online plugin*.



Choose MLST scheme
Select a scheme for an organism from the on-line list or define your own MLST scheme.

● Select organism from on-line list

○ Define your own MLST scheme

Figure 2.3: The *Choose MLST scheme* dialog box.

2.3 Online MLST repositories

In the *Choose MLST scheme* dialog box choose the option ***Select organism from on-line list*** and press <***Next***> to call the *Select an organism* dialog box (see Figure 2.4).

Any organism for which an MLST repository is available online, is listed in the *Select an organism* dialog box. For the selected organism, abbreviations of the housekeeping genes are listed in the right panel, together with the URL where the database is located and the number of available profiles.

Pressing <***Next***> calls the *Define profile and allele files* dialog box (see Figure 2.5).

In the *Define profile and allele files* dialog box, the location of the profile and allele definition files are shown. For public MLST schemes, the default locations point to the correct file URLs.

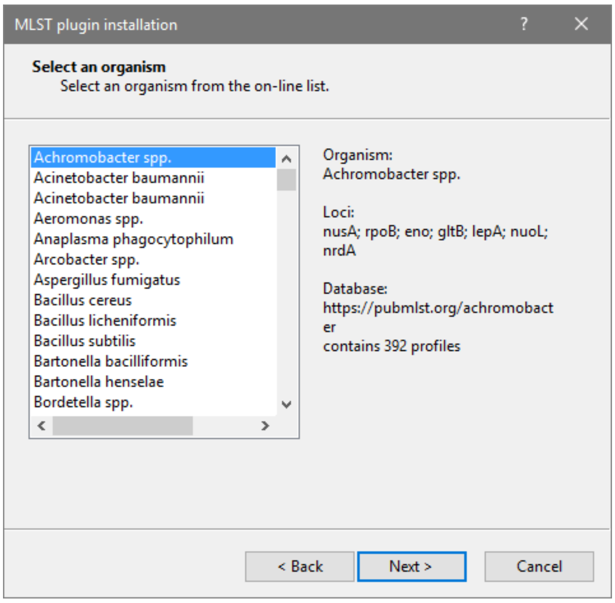
It is possible to specify a different location for the profile and allele definition files, which can be located on your own computer, local area network or on the internet. This is however unnecessary in a standard setup.

Checking the option ***Update profiles and alleles at database startup*** ensures that the profile and allele definitions are always up-to-date, by updating the latter each time when the BioNumerics database is opened.

All remotely stored MLST information (allele numbering, sequence types and clonal complexes) for the selected organism will be downloaded and stored in the BioNumerics database during plugin installation.

Pressing <***Next***> calls the *Allele trimming patterns* dialog box (see Figure 2.6).

In the *Allele trimming patterns* dialog box the start and stop trimming patterns can be entered manually for



Select an organism
Select an organism from the on-line list.

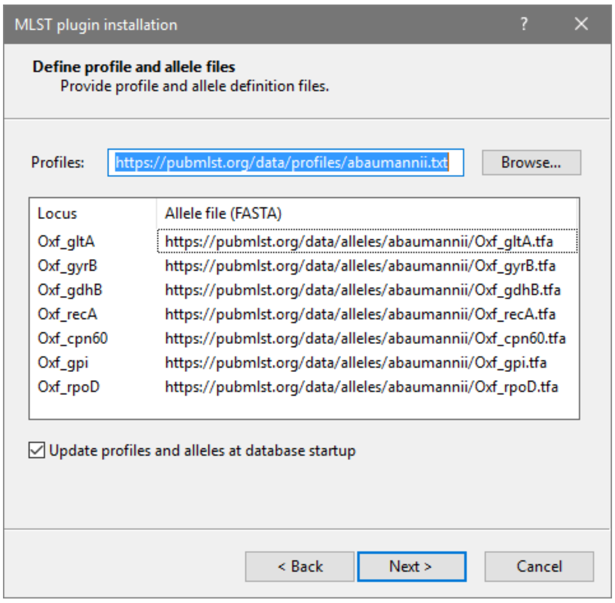
Achromobacter spp.
Acinetobacter baumannii
Acinetobacter baumannii
Aeromonas spp.
Anaplasma phagocytophilum
Arcobacter spp.
Aspergillus fumigatus
Bacillus cereus
Bacillus licheniformis
Bacillus subtilis
Bartonella bacilliformis
Bartonella henselae
Bordetella spp.

Organism:
Achromobacter spp.

Loci:
nusA; rpoB; eno; gltB; lepA; nuoL;
nrdA

Database:
https://pubmlst.org/achromobact
er
contains 392 profiles

Figure 2.4: The *Select an organism* dialog box.



Define profile and allele files
Provide profile and allele definition files.

Profiles: https://pubmlst.org/data/profiles/abaumannii.txt

Locus	Allele file (FASTA)
Oxf_gltA	https://pubmlst.org/data/alleles/abaumannii/Oxf_gltA.tfa
Oxf_gyrB	https://pubmlst.org/data/alleles/abaumannii/Oxf_gyrB.tfa
Oxf_gdhB	https://pubmlst.org/data/alleles/abaumannii/Oxf_gdhB.tfa
Oxf_recA	https://pubmlst.org/data/alleles/abaumannii/Oxf_recA.tfa
Oxf_cpn60	https://pubmlst.org/data/alleles/abaumannii/Oxf_cpn60.tfa
Oxf_gpi	https://pubmlst.org/data/alleles/abaumannii/Oxf_gpi.tfa
Oxf_rpoD	https://pubmlst.org/data/alleles/abaumannii/Oxf_rpoD.tfa

☑ Update profiles and alleles at database startup

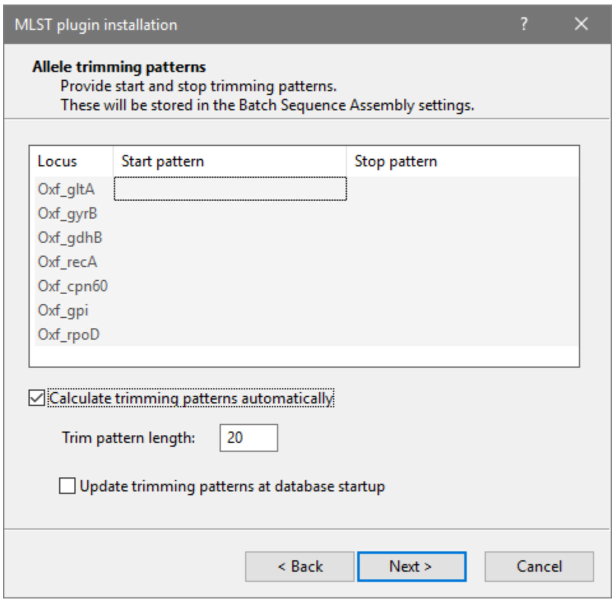
Figure 2.5: The *Define profile and allele files* dialog box.

each of the alleles, or one can specify to ***Calculate trimming patterns automatically*** (recommended). If the trimming patterns are automatically calculated, a ***Trim pattern length*** (default 20 bases) can be set. The trim patterns can be recalculated automatically each time the database is loaded by checking ***Update trimming patterns at database startup***.

Pressing <***Next***> calls the *Database info fields* dialog box (see Figure 2.7).

In the *Database info fields* dialog box, the program prompts for database information fields to store the ***Sequence types*** and ***Clonal complexes*** information. One can either choose the default fields, select existing fields or enter new field names. When manually entering new field names, please note that information field names cannot start with a space.

Pressing <***Next***> calls the final dialog in the wizard (see Figure 2.8).



MLST plugin installation

Allele trimming patterns

Provide start and stop trimming patterns.
These will be stored in the Batch Sequence Assembly settings.

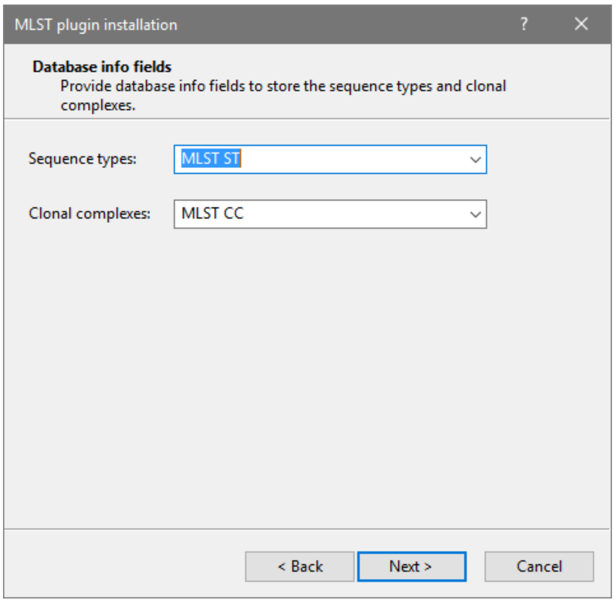
Locus	Start pattern	Stop pattern
Oxf_gltA		
Oxf_gyrB		
Oxf_gdhB		
Oxf_recA		
Oxf_cpn60		
Oxf_gpi		
Oxf_rpoD		

☑ Calculate trimming patterns automatically

Trim pattern length: 20

☐ Update trimming patterns at database startup

Figure 2.6: The *Allele trimming patterns* dialog box.



MLST plugin installation

Database info fields

Provide database info fields to store the sequence types and clonal complexes.

Sequence types: MLST ST

Clonal complexes: MLST CC

Figure 2.7: The *Database info fields* dialog box.

Pressing <***Finish***> in the *Finish installation* dialog box starts with the installation of the *MLST online plugin*.

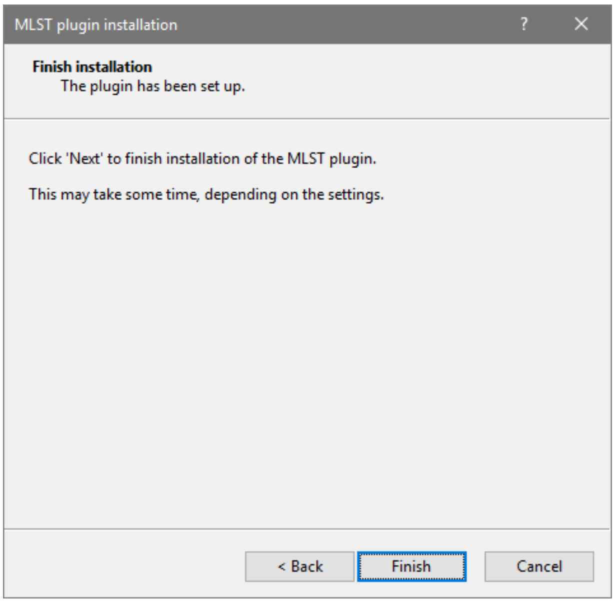
In case of large MLST databases available online, the installation may take several minutes, depending on the speed of your internet connection.

When the *MLST online plugin* is successfully installed, a confirmation message is displayed. Close and reopen the database to activate the features of the *MLST online plugin*.

The *MLST online plugin* installs menu items in the main menu of the software under ***MLST*** (see Figure 2.9) and in the *Contig assembly* window.

In the *Main* window, the *MLST online plugin* has installed following items:

- Extra information fields in the *Database entries* panel (default names: **MLST ST**; **MLST CC**).



Finish installation

The plugin has been set up.

Click 'Next' to finish installation of the MLST plugin.

This may take some time, depending on the settings.

Figure 2.8: The *Finish installation* dialog box.

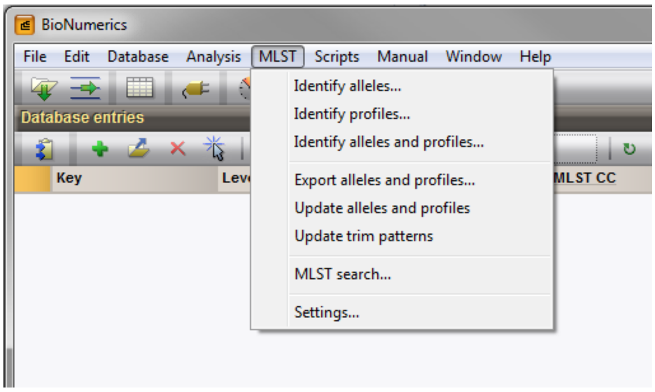


Figure 2.9: MLST menu items in the *Main* window.

- One character type called **MLST**, one composite dataset called **MLST_CMP**, and typically seven sequence types, each named after a housekeeping gene.

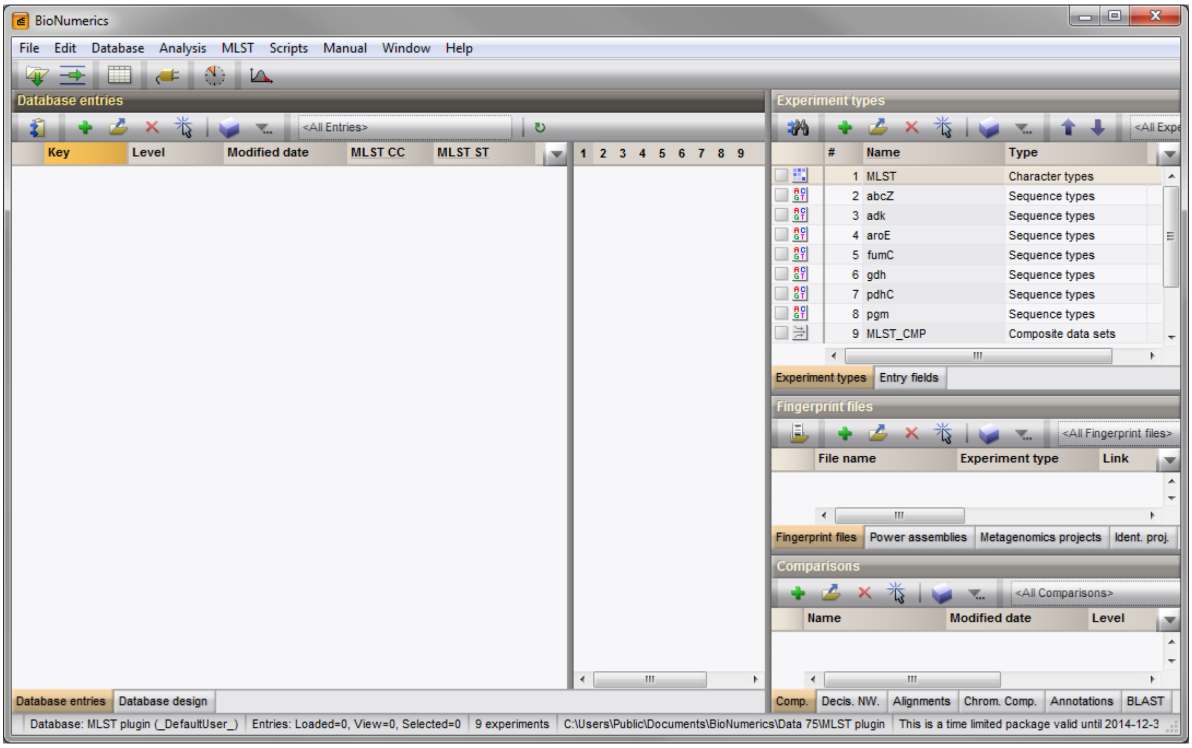
2.4 Creating a custom MLST scheme

In case no MLST scheme is available yet for the organism of your interest, the *MLST online plugin* allows you to create a custom MLST scheme.

In the *Choose MLST scheme* dialog box (see Figure 2.3), select ***Define your own MLST scheme*** and press <***Next***> to call the *Define your own MLST scheme* dialog box (see Figure 2.11).

In the *Define your own MLST scheme* dialog box, you are asked to provide the ***Organism name*** and the ***Loci*** that were sequenced. The loci should be entered, separated by semi-colons.

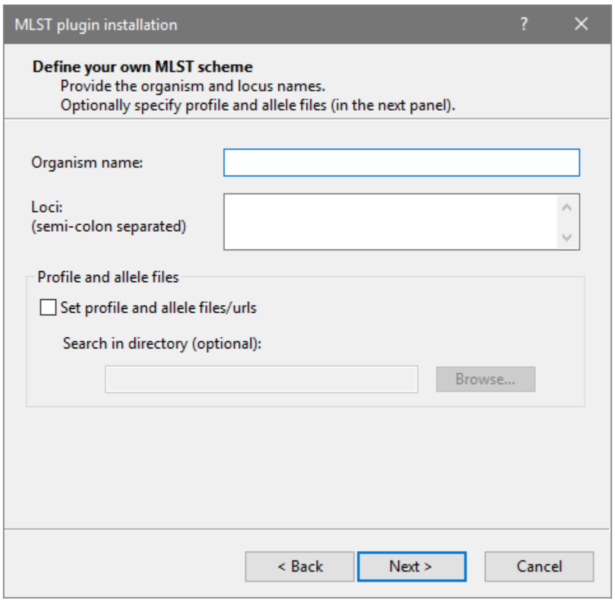
In case text files containing profile and allele information are already available (e.g. if they were sent to you by a colleague who works on the same organism), you should check ***Set profile and allele files/URLs***. With this option checked, pressing <***Next***> will take you to the *Select an organism* dialog box where profile and allele file locations can be specified (see Figure 2.5), which will initially be empty.


BioNumerics
File Edit Database Analysis MLST Scripts Manual Window Help
Database entries
<All Entries>
Key Level Modified date MLST CC MLST ST
1 2 3 4 5 6 7 8 9
Experiment types
<All Expe
Experiment types Entry fields
Fingerprint files
<All Fingerprint files>
File name Experiment type Link
Fingerprint files Power assemblies Metagenomics projects Ident. proj.
Comparisons
<All Comparisons>
Name Modified date Level
Database entries Database design
Comp. Decis. NW. Alignments Chrom. Comp. Annotations BLAST

#	Name	Type
1	MLST	Character types
2	abcZ	Sequence types
3	adk	Sequence types
4	aroE	Sequence types
5	fumC	Sequence types
6	gdh	Sequence types
7	pdhC	Sequence types
8	pgm	Sequence types
9	MLST_CMP	Composite data sets

Database: MLST plugin (_DefaultUser_) | Entries: Loaded=0, View=0, Selected=0 | 9 experiments | C:\Users\Public\Documents\BioNumerics\Data 75\MLST plugin | This is a time limited package valid until 2014-12-3

Figure 2.10: The *Main* window after installation of the *MLST online plugin*.



Define your own MLST scheme
Provide the organism and locus names.
Optionally specify profile and allele files (in the next panel).

☐ Set profile and allele files/urls

Figure 2.11: The *Define your own MLST scheme* dialog box.

When ***Set profile and allele files/URLs*** is unchecked, the next step will be dealing with the trimming positions.

When FASTA text files for allele definitions are used, they should be formatted as follows:

```
> [ID]
ACTG...
```

With [ID] an integer allele identifier (without brackets). Alternatively, the identifier line could also be:

”>`_[ID]` ”, ”>`-[ID]` ”, ”>`[LOCUS] [ID]` ”, ”>`[LOCUS]_[ID]` ” or ”>`[LOCUS]-[ID]`.

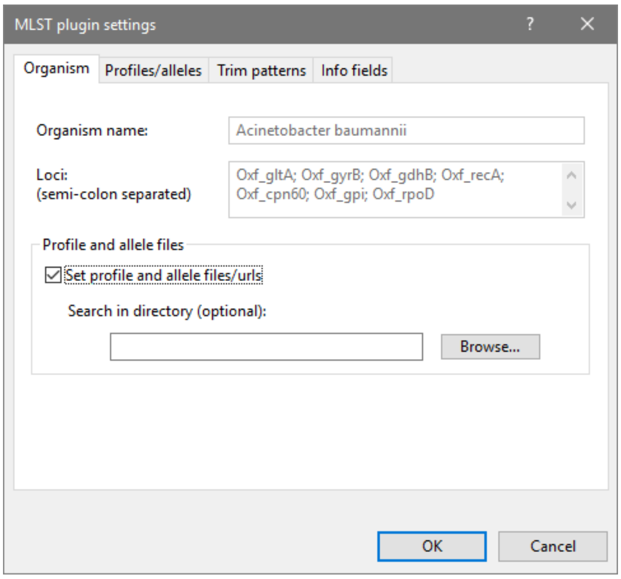
When the option ***Set profile and allele files/URLs*** was checked, specify the profile and allele file locations in the *Select an organism* dialog box.

When the option ***Set profile and allele files/URLs*** was unchecked, enter the ***Start pattern*** and ***Stop pattern*** for all loci or check ***Calculate trimming patterns automatically*** in the *Allele trimming patterns* dialog box.

Proceed with the steps explained in 2.3 to complete the installation of the *MLST online plugin*.

2.5 MLST plugin settings

All settings that were entered during installation of the MLST online plugin, except organism name and loci, can be changed later on. In the *Main* window, select ***MLST*** > ***Settings*** to call the *MLST plugin settings* dialog box (see Figure 2.12).



Organism name: Acinetobacter baumannii

Loci:
(semi-colon separated) Oxf_gltA; Oxf_gyrB; Oxf_gdhB; Oxf_recA; Oxf_cpn60; Oxf_gpi; Oxf_rpoD

☑ Set profile and allele files/urls

Search in directory (optional):

Figure 2.12: The *MLST plugin settings* dialog box, *Organism tab*.

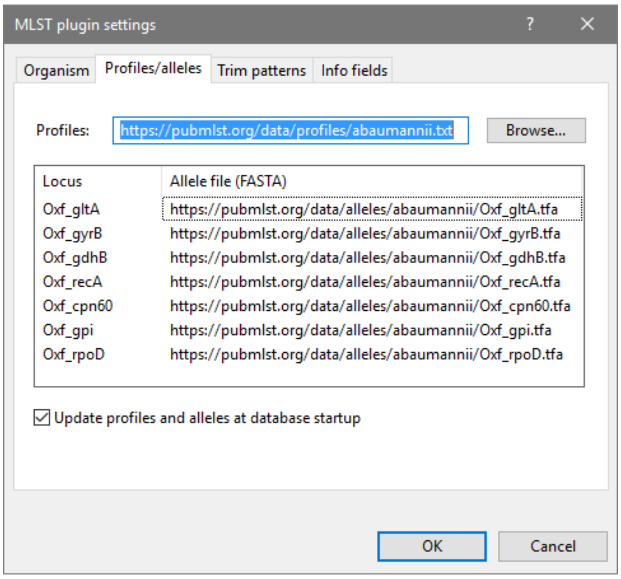
The *MLST plugin settings* dialog box consists of four different tabs: the *Organism*, *Profiles/alleles*, *Trim patterns*, and *Info fields tab*. To access the relevant settings, simply click on the corresponding tab in the top part of the dialog box.

In the *Organism tab*, which is displayed initially, the ***Organism name*** and ***Loci*** are read-only, since the MLST scheme cannot be changed after installation. Checking ***Set profile and allele files/URLs*** will update profile and allele definitions from external text files, which can be located on your computer, on a network computer or on the internet. Optionally, a directory can be entered or browsed for.

The *Profiles/alleles tab* basically contains the same settings as the *Define profile and allele files* dialog box (see Figure 2.5). The location where the program looks for the profile and/or allele files can be edited. This action is of course only necessary when the physical location of these files on your computer or on the (web) server has changed.

In the *Trim patterns tab*, the ***Start pattern*** and ***Stop pattern*** for each of the loci can be edited (only if ***Calculate trimming patterns automatically*** is unchecked), similar as in the *Allele trimming patterns* dialog box (see Figure 2.6). Also, it can be specified to update the trimming patterns each time that the database is loaded. A ***Trim pattern length*** can be entered if the trimming patterns are calculated automatically.

From the *Info fields tab*, different information fields can be selected for storage of ***Sequence types*** and

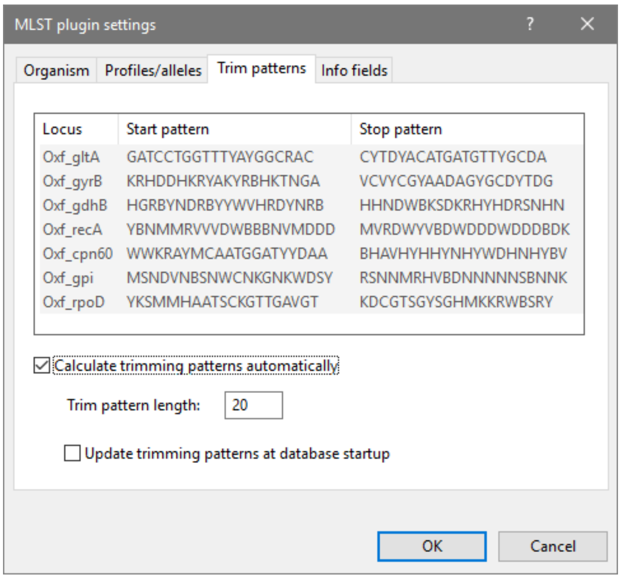


Profiles: https://pubmlst.org/data/profiles/abaumannii.txt

Locus	Allele file (FASTA)
Oxf_gltA	https://pubmlst.org/data/alleles/abaumannii/Oxf_gltA.tfa
Oxf_gyrB	https://pubmlst.org/data/alleles/abaumannii/Oxf_gyrB.tfa
Oxf_gdhB	https://pubmlst.org/data/alleles/abaumannii/Oxf_gdhB.tfa
Oxf_recA	https://pubmlst.org/data/alleles/abaumannii/Oxf_recA.tfa
Oxf_cpn60	https://pubmlst.org/data/alleles/abaumannii/Oxf_cpn60.tfa
Oxf_gpi	https://pubmlst.org/data/alleles/abaumannii/Oxf_gpi.tfa
Oxf_rpoD	https://pubmlst.org/data/alleles/abaumannii/Oxf_rpoD.tfa

☑ Update profiles and alleles at database startup

Figure 2.13: The *MLST plugin settings* dialog box, *Profiles/alleles tab*.



Locus	Start pattern	Stop pattern
Oxf_gltA	GATCCTGGTTTYAYGGCRAC	CYTDYACATGATGTTYGCDA
Oxf_gyrB	KRHDDHKRYAKYRBHKTNGA	VCVYCGYAADAGYGCDYTDG
Oxf_gdhB	HGRBYNDRBYYWVHRDYNRB	HHNDWBKSDKRHYHDRSNHN
Oxf_recA	YBNMMRVVVDWBBBNVMDDD	MVRDWYVBDWDDDWDDDBDK
Oxf_cpn60	WWKRAYMCAATGGATYYDAA	BHAVHYHHYNHYWDHNHYBV
Oxf_gpi	MSNDVNBSNWCNKGNKWDSY	RSNNMRHVBDNNNNNSBNNK
Oxf_rpoD	YKSMMHAATSCKGTTGAVGT	KDCGTSGYSGHMKKRWBSRY

☑ Calculate trimming patterns automatically

Trim pattern length: 20

☐ Update trimming patterns at database startup

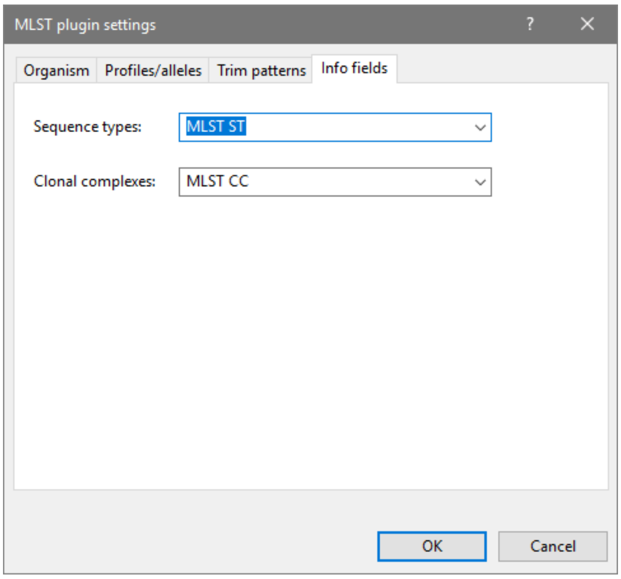
Figure 2.14: The *MLST plugin settings* dialog box, *Trim patterns tab*.

Clonal complexes information. The functionality of this tab corresponds to the *Database info fields* dialog box (see Figure 2.7).

2.6 Updating the allele trimming patterns

All trimming patterns are displayed in the *Trim patterns tab* of the *MLST plugin settings* dialog box (see 2.5). In the *Trim patterns tab*, the ***Start*** and ***Stop patterns*** for each of the loci are displayed. These patterns will be automatically displayed in the *Assembly trimming settings* dialog box when importing sequences into the database with the batch import routine.

When the option ***Update trimming patterns at database startup*** is checked in the *Trim patterns tab*, the trim patterns are automatically updated each time the database is loaded.



Sequence types: MLST ST

Clonal complexes: MLST CC

Figure 2.15: The The *MLST plugin settings* dialog box, *Info fields tab*.

To update the trim patterns of the selected organism manually, use the ***MLST*** > ***Update trim patterns*** command.

Chapter 3

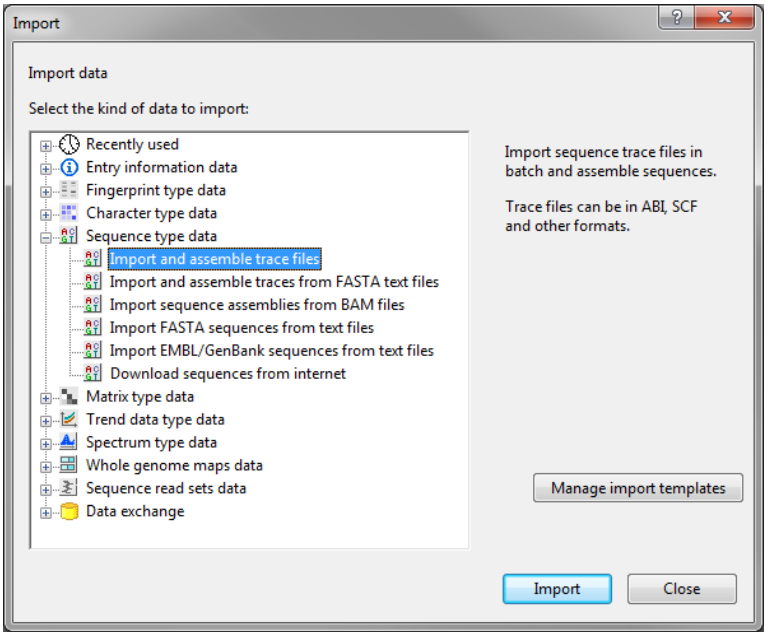
Importing sequences and sequence trace files

3.1 Introduction

Sequences and sequence trace files can be imported (and assembled) in BioNumerics and stored in the corresponding housekeeping gene sequence type experiments using the import routines available in the *Import* dialog box. After import, the (consensus) sequences can be screened against the downloaded MLST sequences (see 5.1 for more information).

3.2 Import routines

The *Import* dialog box is called with the command ***File*** > ***Import...*** (⬇, **Ctrl+I**). The import tree options are organized in groups based upon the type of data. Sequence data can be imported in BioNumerics in several ways (see Figure 3.1).


Import
Import data
Select the kind of data to import:
Recently used
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
Whole genome maps data
Sequence read sets data
Data exchange
Manage import templates
Import
Close

Import sequence trace files in batch and assemble sequences.

Trace files can be in ABI, SCF and other formats.

Figure 3.1: The sequence import routines in the Import tree.

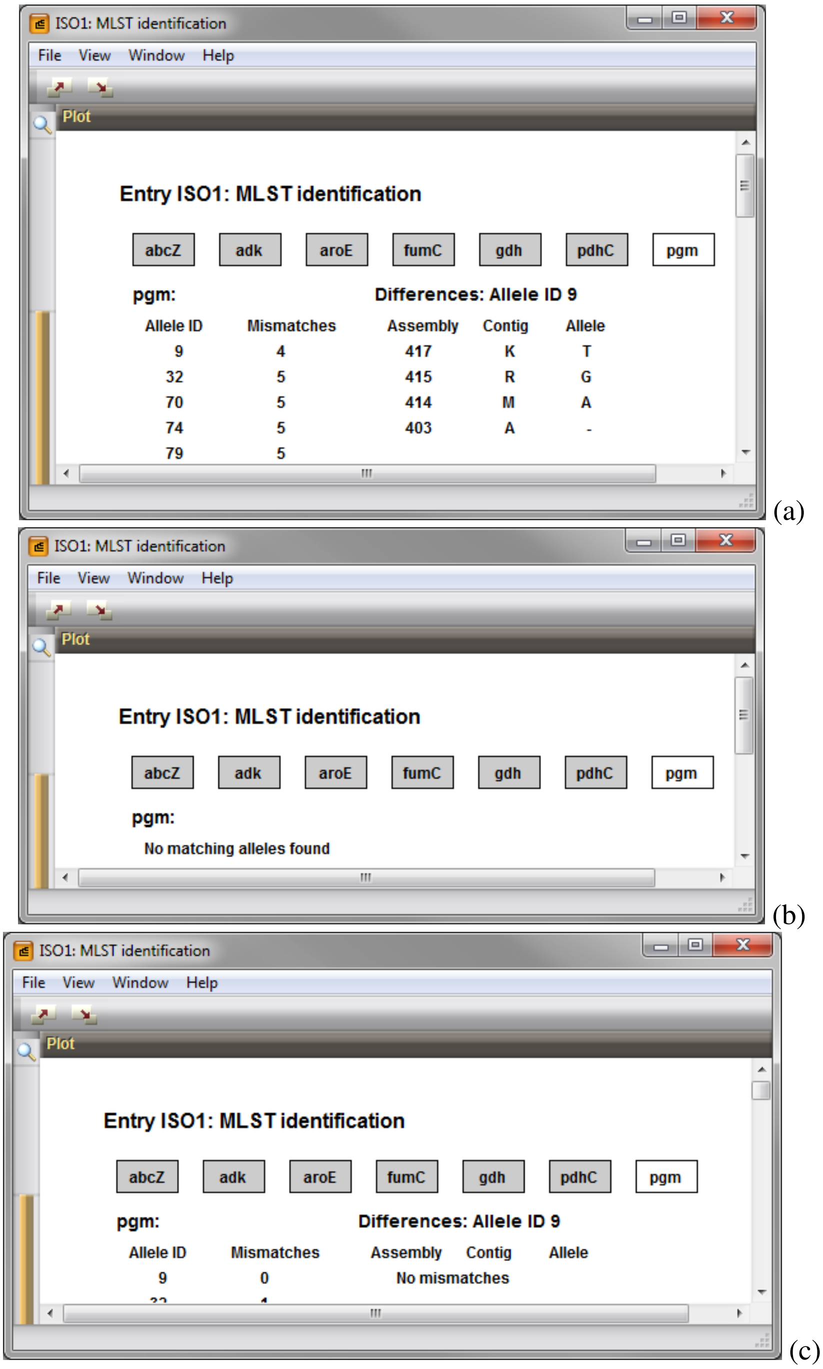
The import routines will not be covered in detail in this manual. More detailed information can be found in the BioNumerics manual.

3.3 The Assembler window

When importing and assembling sequence trace files in BioNumerics, the consensus sequence can be screened for (nearest) allele matches with ***MLST*** > ***Identify allele*** in the *Contig assembly* window.

The *MLST identification plot window* will pop up. The sequence type of the sequence that is shown in Assembler is displayed in white. The other sequence types are shown in gray (see Figure 3.2). The identification result is shown below the sequence type boxes.

In case the trimmed consensus sequence could not be matched to one of the alleles (e.g. due to the presence of unresolved bases), a table is shown in the lower left part of the *MLST identification plot window* displaying the best matched alleles in the ***Allele ID*** column and number of mismatches with the consensus sequence in the ***Mismatches*** column (see Figure 3.2 a). Matched allele IDs with a maximum of 5 mismatches are reported. All allele IDs are clickable. A second table is displayed in the right part of the window listing the sequence differences between the consensus sequence and the selected allele ID in the left table. The base and base position on the consensus sequence are displayed in the ***Contig*** and ***Assembly*** columns, the corresponding bases on the matched allele are displayed in the ***Allele*** column. Clicking on one of the editing suggestions in the right table automatically updates the focus in the *Contig assembly* window. In case no match could be found after introduction of 5 or less mismatches, the text ”No matching alleles found” is displayed (see Figure 3.2 b).



Entry ISO1: MLST identification

pgm:

Allele ID	Mismatches
9	4
32	5
70	5
74	5
79	5

Differences: Allele ID 9

Assembly	Contig	Allele
417	K	T
415	R	G
414	M	A
403	A	-

(a)

Entry ISO1: MLST identification

pgm:

No matching alleles found

(b)

Entry ISO1: MLST identification

pgm:

Allele ID	Mismatches
9	0

Differences: Allele ID 9

Assembly	Contig	Allele

No mismatches

(c)

Figure 3.2: The *MLST identification plot window*: (a) Best match with allele ID 9; (b) No matching alleles; (c) Perfect match with allele ID 9.

Chapter 4

Importing allele numbers from external files

4.1 Introduction

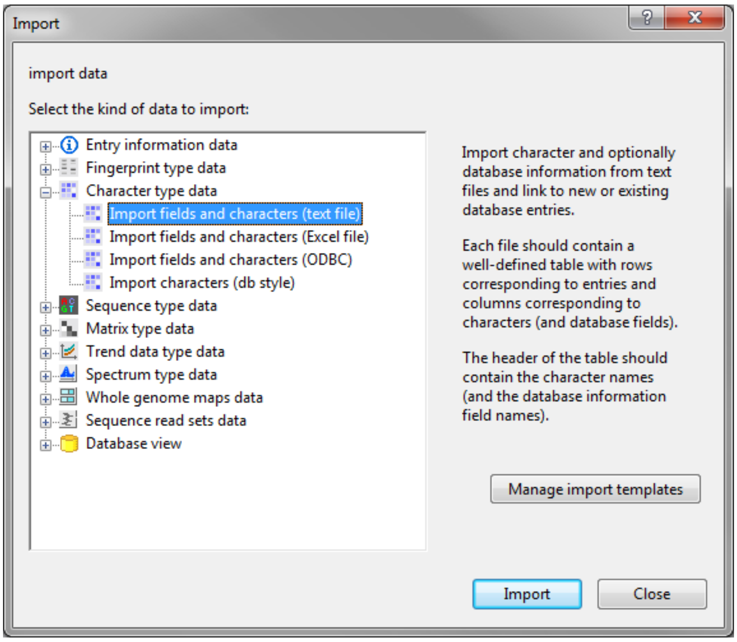
When the MLST allele numbers are already stored in external text, Excel, or other ODBC-compatible file, these allele numbers can be imported in BioNumerics and stored in the character type experiment **MLST** using import routines available in the *Import* dialog box. After import, the allelic profiles can be screened against the sequence type and clonal complex information (see 5.2 for more information).

4.2 Import routines

The *Import* dialog box is called with the command ***File*** > ***Import...*** (⬇, **Ctrl+I**). The import tree options are organized in groups based upon the type of data (see Figure 4.1).

- With the ***Import fields and characters (text file)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from text files in the database and linked to new or existing database entries.
- With the ***Import fields and characters (Excel file)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from an Excel file in the database and linked to new or existing database entries.
- With the ***Import fields and characters (ODBC)*** option, listed under the topic ***Character type data*** in the Import tree, the allele numbers can be imported from ODBC-compatible files in the database and linked to new or existing database entries.

The import routines will not be covered in detail in this manual. More detailed information can be found in the BioNumerics manual.



Import character and optionally database information from text files and link to new or existing database entries.

Each file should contain a well-defined table with rows corresponding to entries and columns corresponding to characters (and database fields).

The header of the table should contain the character names (and the database information field names).

Figure 4.1: The character import routines in the Import tree.

Chapter 5

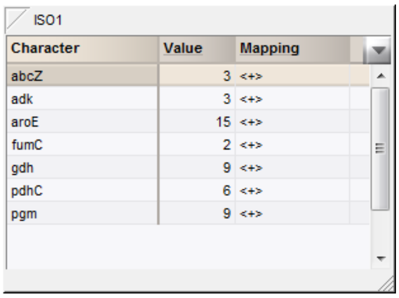
Alleles and profiles

5.1 Identifying alleles

After clean-up of the assemblies, the consensus sequences can be screened against the allele information with ***MLST* > *Identify alleles***.

Screening can be done for all entries present in the database, or for any selection of entries in database. A selection can be made in the database using the **Ctrl** and **Shift** keys. To select all entries in the database at once, use the shortcut **Ctrl+A**. Check boxes for selected entries are indicated as ☑. If no selection is present in the database, a message pops up prompting to confirm to run the tool on all entries in the database. Pressing <***OK***> starts the allele identification for all entries that are present in the database.

The matched allele IDs are stored in the corresponding character fields of the **MLST** character type. When a consensus sequence does not match one of the alleles in the database, the character value is left empty. Clicking on the colored dot in the **MLST** column of the *Experiment presence* panel opens the character *Experiment card* window for an entry (see Figure 5.1). The experiment card can be closed again by clicking in the small triangle-shaped button in the left upper corner.



ISO1

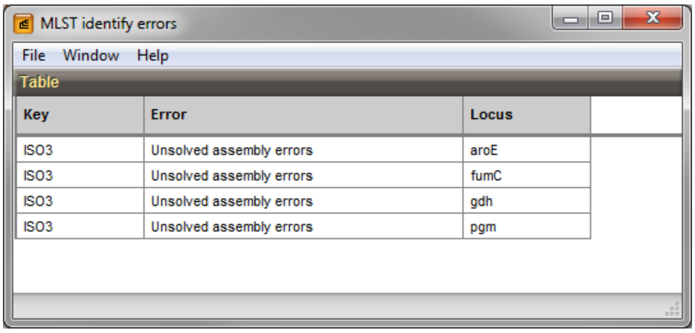
Character	Value	Mapping
abcZ	3	<+>
adk	3	<+>
aroE	15	<+>
fumC	2	<+>
gdh	9	<+>
pdhC	6	<+>
pgm	9	<+>

Figure 5.1: An MLST character card, displaying allele IDs.

If one or more sequences could not be assigned to an allele ID, an *MLST error report window* pops up (see Figure 5.2 for an example). The report window displays all errors encountered during the search for allele IDs ('Error' column), the linked sequence type ('Locus' column) and the entry key ('Key' column).

The error messages can be:

- ***Unresolved assembly errors***: When the status of the assembly is either "error" or "read", i.e. when the assembly errors were not yet manually corrected or when the errors have been corrected but the status has not been changed to "solved" afterwards.
- ***No sequence match***: The consensus sequence could not be matched to an existing allele ID.

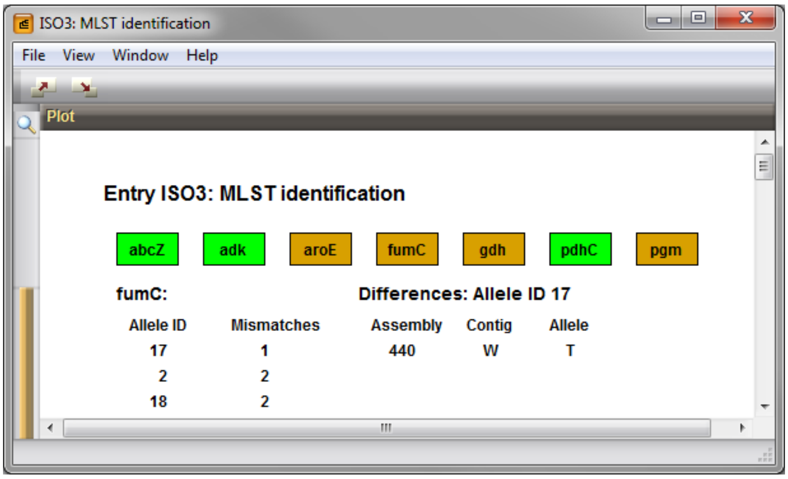


Key	Error	Locus
ISO3	Unsolved assembly errors	aroE
ISO3	Unsolved assembly errors	fumC
ISO3	Unsolved assembly errors	gdh
ISO3	Unsolved assembly errors	pgm

Figure 5.2: The *MLST error report window*, after identifying alleles.

Double-clicking on one of the error messages in the *MLST error report window* opens the *MLST identification plot window* of the linked entry (see Figure 5.3).

The sequence types that are assigned to an allele ID are displayed in green. The sequence types of the sequences that could not be linked to an allele ID are displayed in orange (see Figure 5.3).



Entry ISO3: MLST identification

fumC:

Allele ID	Mismatches
17	1
2	2
18	2

Differences: Allele ID 17

Assembly	Contig	Allele
440	W	T

Figure 5.3: The *MLST identification plot window*: No perfect match with an allele, best match with allele ID 17.

The identification result of the selected sequence type is shown below the sequence plot:

In case the trimmed consensus sequence could not be matched to one of the alleles (e.g. due to the presence of unresolved bases) a table is shown in the lower left part of the *MLST identification plot window* displaying the best matched allele IDs in the ”Allele ID” column, and number of mismatches with the consensus sequence in the ”Mismatches” column (see Figure 5.3). Matched allele IDs with a maximum of 5 mismatches are reported. All allele IDs can be clicked on, in which case a second table is displayed in the right part of the window, displaying the sequence differences between the consensus sequence and the selected allele ID in the left table. The base and base position on the consensus sequence are displayed in the ”Contig” and ”Assembly” columns respectively. The corresponding bases on the matched allele are displayed in the ”Allele” column. Clicking on one of the sequence differences in the right table, automatically updates the focus in the *Contig assembly* window. In case no match could be found after introduction of 5 or less mismatches, the text ”No matching alleles found” is displayed.

Clicking on another sequence type field automatically updates the identification results in the lower part of the *MLST identification plot window*.

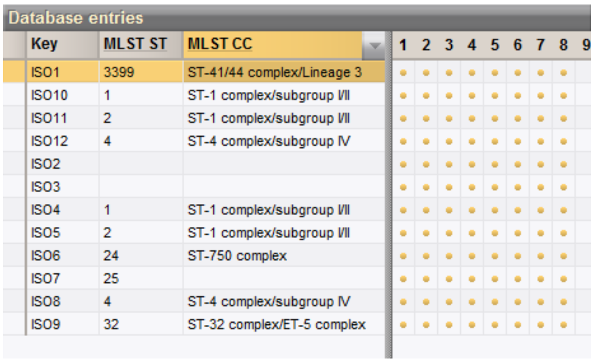
Clean up all assembly errors in the *MLST error report window* and select ***MLST*** > ***Identify alleles*** to update the allelic profiles for the selected entries in the database.

5.2 Identifying allelic profiles

The allelic profiles of the entries in the database are screened against the downloaded sequence type and clonal complex information with the command ***MLST*** > ***Identify profiles***.

Screening can be done for all entries present in the database, or for any selection of entries in database. A selection can be made in the database using the **Ctrl-** and **Shift**-keys. To select all entries in the database at once, use the shortcut **Ctrl+A**. Check boxes for selected entries are indicated as ☑. If no selection is present in the database, a message pops up prompting to confirm to run the tool on all entries in the database. Pressing <***OK***> starts the profile screening for all entries in the database.

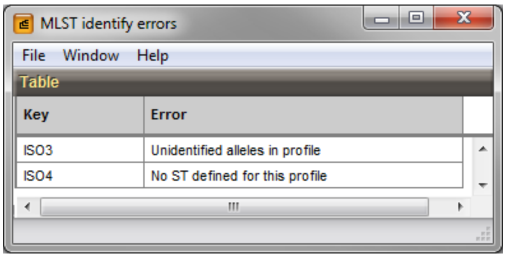
The matched sequences types and clonal complexes are displayed in the MLST information fields (default ”MLST ST” and ”MLST CC” respectively, see Figure 5.4).



Key	MLST ST	MLST CC	1	2	3	4	5	6	7	8	9
ISO1	3399	ST-41/44 complex/Lineage 3	●	●	●	●	●	●	●	●	
ISO10	1	ST-1 complex/subgroup I/II	●	●	●	●	●	●	●	●	
ISO11	2	ST-1 complex/subgroup I/II	●	●	●	●	●	●	●	●	
ISO12	4	ST-4 complex/subgroup IV	●	●	●	●	●	●	●	●	
ISO2			●	●	●	●	●	●	●	●	
ISO3			●	●	●	●	●	●	●	●	
ISO4	1	ST-1 complex/subgroup I/II	●	●	●	●	●	●	●	●	
ISO5	2	ST-1 complex/subgroup I/II	●	●	●	●	●	●	●	●	
ISO6	24	ST-750 complex	●	●	●	●	●	●	●	●	
ISO7	25		●	●	●	●	●	●	●	●	
ISO8	4	ST-4 complex/subgroup IV	●	●	●	●	●	●	●	●	
ISO9	32	ST-32 complex/ET-5 complex	●	●	●	●	●	●	●	●	

Figure 5.4: The *Main* window after screening for sequence type and clonal complex information.

If one or more allelic profiles could not be assigned to a sequence type (or clonal complex), an *MLST error report window* pops up (see Figure 5.5 for an example). The report window displays the errors encountered during the search for sequence types and clonal complexes. The errors are shown in the ’Error’ column, and the linked entry key is displayed in the ’Key’ column.



Key	Error
ISO3	Unidentified alleles in profile
ISO4	No ST defined for this profile

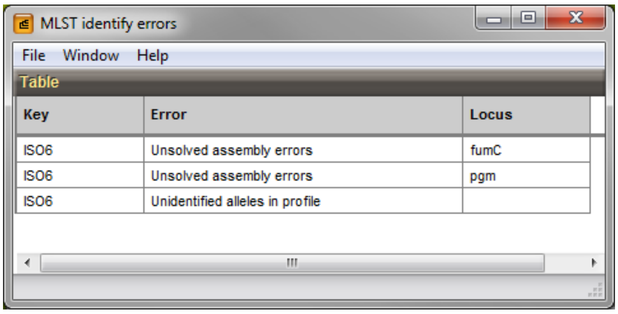
Figure 5.5: The *MLST error report window*, after identifying profiles.

The error messages can be:

- ***Unidentified alleles in profile***: One or more alleles are not assigned to an allele ID.
- ***No sequence type match***: The allelic profile could not be matched to an existing allelic profile in the locally stored database.
- ***No ST defined for this profile***: The allelic profile could not be matched to an existing allelic profile in the online database.
- ***No complex defined for this ST***: The sequence type is not linked to a clonal complex in the online database.

5.3 Identifying alleles and allelic profiles

A combined screening for alleles and profiles can be launched with the command ***MLST*** > ***Identify alleles and profiles***. This command combines the actions ***MLST*** > ***Identify allele*** and ***MLST*** > ***Identify profiles*** as described in 5.1 and 5.2, respectively. If one or more sequences could not be assigned to an allele ID and/or if one or more allelic profiles could not be assigned to a sequence type (or clonal complex), an *MLST error report window* pops up, which displays all errors encountered during the search for allele IDs, sequence types, and clonal complexes (see Figure 5.6 for an example).



Key	Error	Locus
ISO6	Unsolved assembly errors	fumC
ISO6	Unsolved assembly errors	pgm
ISO6	Unidentified alleles in profile	

Figure 5.6: The *MLST error report window*, after identifying alleles and profiles.

5.4 Updating locally stored allele and profile information

Locally stored information is used when querying for allele, sequence type and clonal complex information in BioNumerics. This information can be obtained from profile and allele definition files, located on your computer, local area network or on the internet (see Figure 2.5). With the command ***MLST*** > ***Update alleles and profiles***, the locally stored MLST information in the BioNumerics database is then compared with the information in the allele and profile files and updated when new information is found. A warning is generated if existing alleles or profiles are changed or deleted. When the option ***Update profiles and alleles at database startup*** is checked, the MLST information is automatically updated each time the database is loaded. In case profile and allele files are not available, the locally stored MLST information should be manually updated (see 5.5).

5.5 Editing MLST information in a custom MLST database

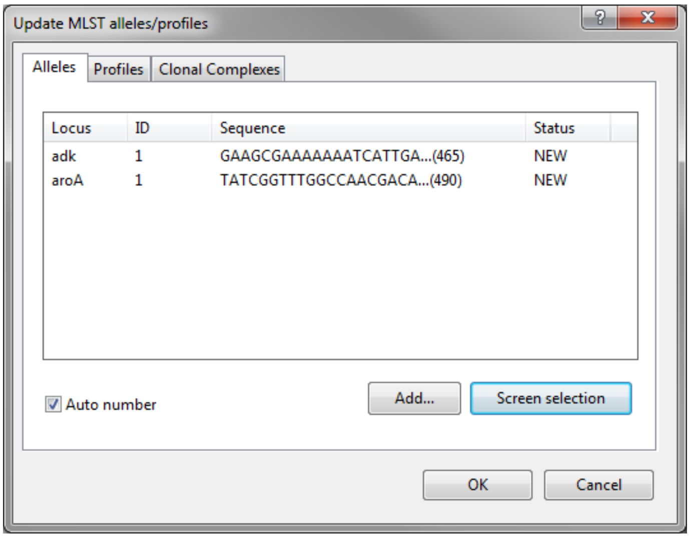
5.5.1 Introduction

The *locally* stored MLST information of a *custom* MLST database can be updated from profile and allele definition files specified in the *Profiles/alleles tab* (see 5.4).

If no allele and profile definition files are specified, the MLST information can be updated manually using the *Update MLST alleles/profiles* dialog box. Selecting ***MLST*** > ***Update alleles and profiles*** calls this dialog (see Figure 5.7). The *Update MLST alleles/profiles* dialog box consists of three different tabs: the *Alleles*, *Profiles* and *Clonal Complexes tab*. To access the relevant settings, simply click on the corresponding tab in the top part of the dialog box.

5.5.2 Editing alleles

In the *Alleles tab*, sequences can be added to the database.



Locus	ID	Sequence	Status
adk	1	GAAGCGAAAAAAATCATTGA...(465)	NEW
aroA	1	TATCGGTTTGGCCAACGACA...(490)	NEW

☑ Auto number

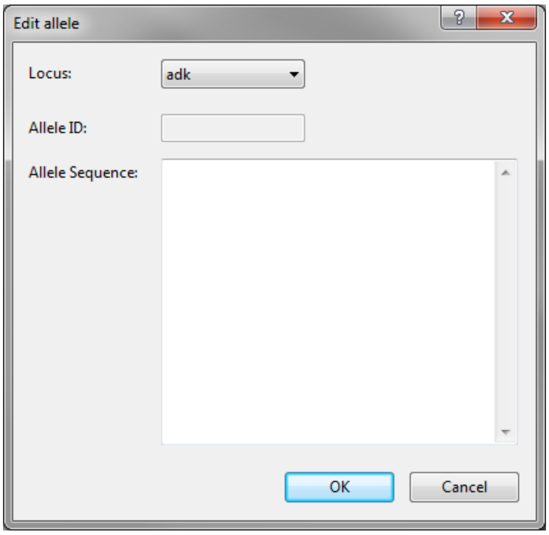
Figure 5.7: Allele information in the custom MLST database.

When pressing the <***Screen selection***> button, the software compares all sequences of the current entry selection with the sequences that are already present in the custom database.

All newly detected sequences are added to the allele list in the *Alleles tab*. The 'Locus' column displays the name of the sequence type, the 'Sequence' field displays the first nucleotides of the allele sequence, and the last column displays the 'Status' information. If the option ***Auto number*** was unchecked, no IDs are assigned to the newly added sequences in the 'ID' column. An ID can be added to a sequence by double-clicking on a sequence in the list. If the ***Auto number*** option was checked, the software automatically assigns unique IDs to the newly added sequences (recommended).

Double-click on an allele in the list if you want to edit the allele information. This calls the *Edit allele* dialog box.

Instead of letting the program screen a selection of entries for alleles that are not yet present in the database, a new allele can also be added manually by pressing the <***Add***> button. This calls the *Edit allele* dialog box.



Locus: adk

Allele ID:

Allele Sequence:

Figure 5.8: Adding an allele to the custom MLST database.

Select the sequence type from the ***Locus*** drop-down menu and enter the sequence in the ***Allele Sequence*** text box. If the option ***Auto number*** is unchecked in the *Alleles tab*, specify a unique allele number in the ***Allele ID*** text box. If the ***Auto number*** option is checked, the ***Allele ID*** text box is not editable and the software will automatically assign a new ID to the new allele when the allele is added to the database.

Pressing <***OK***> in the *Update MLST alleles/profiles* dialog box adds all new alleles in the list to the database.

It is possible to view the alleles or allelic profiles stored in the database with an *object query*. In the *Main* window, choose ***Database*** > ***Object queries...*** (▦) and select "<Create new>" from the drop-down menu that appears. As ***Object to report***, select "MLST - Alleles" or "MLST - Sequence types" and press <***OK***>.

The update alleles and profiles function furthermore allows you to edit and delete alleles that are already stored in the database. An object query (see above) can be used to retrieve the allele ID of the allele to be edited.

In the *Alleles tab*, uncheck ***Auto number*** and press the <***Add***> button to call the *Edit allele* dialog box (see Figure 5.8). Select the sequence type from the ***Locus*** drop-down menu and enter the ***Allele ID*** of the allele to be edited.

To *edit* the allele, paste a corresponding sequence in the ***Allele Sequence*** text box and press <***OK***>. A confirmation message will appear, confirming the change. If you press <***Yes***>, the edited allele will be listed as "UPDATE" in the allele list. The allele will effectively be edited in the database after pressing the <***OK***> button in the *Update MLST alleles/profiles* dialog box.

To *delete* the allele from the database, leave the ***Allele Sequence*** text box empty and press <***OK***>. A confirmation message will appear, confirming the removal of the allele. If you press <***Yes***>, the allele will be listed as "DELETE" in the allele list. The allele will effectively be deleted from the database after pressing the <***OK***> button in the *Update MLST alleles/profiles* dialog box.

Please note that adding new alleles to the database is also possible from within Assembler. When identifying an allele in the *Contig assembly* window, new alleles might be found, i.e. no matching allele is detected in the database, while no errors or inconsistencies are present in the alignment. In that case, you can select ***MLST*** > ***Add to alleles*** in the *Contig assembly* window to add the allele to the database with an automatically assigned number. When the allele indeed does not yet exist in the database, a confirmation message will appear. An error will be generated in case the allele was already present in the database.

5.5.3 Editing profiles

In the *Profiles tab*, profile information can be added to the database.

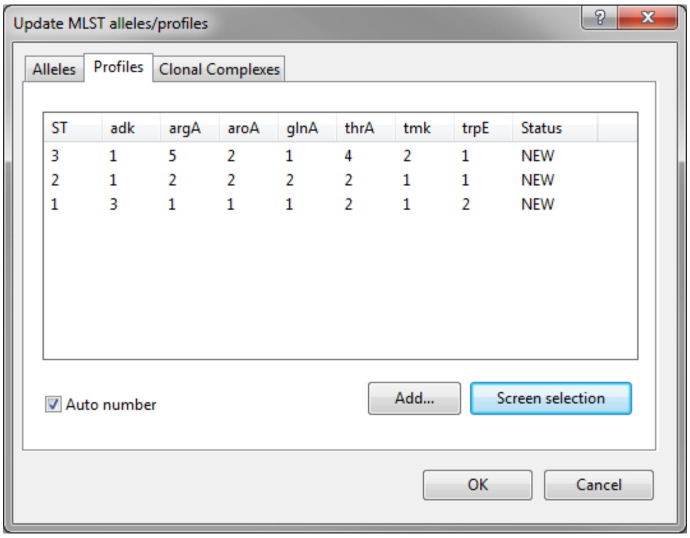
When pressing the <***Screen selection***> button, the software compares the allelic profiles of the current entry selection to the profiles that are already present in the custom database. All newly detected complete profiles are added to the profiles list.

The ***ST*** column displays the name of the sequence type; the last column displays the ***Status*** information. If the option ***Auto number*** was unchecked in the *Profiles tab*, no sequence types are assigned to the newly added profiles in the list. A ST can be added by double-clicking on a profile in the list. If the ***Auto number*** option was checked, the software automatically assigns unique ST IDs to the newly added profiles.

Double-click on a profile in the list if you want to edit the profile information. This pops up the *Edit profile* dialog box.

A new profile can also be added manually by pressing the <***Add***> button. This calls the *Edit profile* dialog box.

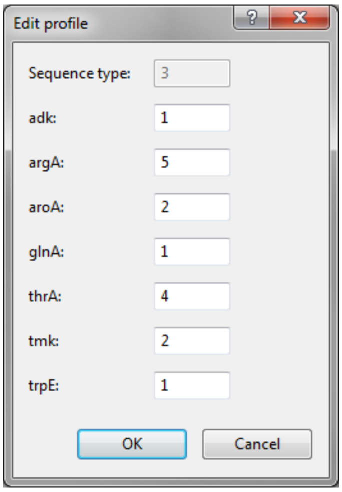
Enter the profile in the text boxes. If the option ***Auto number*** is unchecked in the *Profiles tab*, specify a unique sequence type number in the ***Sequence type*** text box. If the ***Auto number*** option is checked, the ***Sequence type*** text box is not editable and the software will automatically assign a new ID to the new profile



ST	adk	argA	aroA	glnA	thrA	tmk	trpE	Status
3	1	5	2	1	4	2	1	NEW
2	1	2	2	2	2	1	1	NEW
1	3	1	1	1	2	1	2	NEW

☑ Auto number

Figure 5.9: Profile information in the custom MLST database.



Sequence type: 3
adk: 1
argA: 5
aroA: 2
glnA: 1
thrA: 4
tmk: 2
trpE: 1

Figure 5.10: Adding an allelic profile to the custom MLST database.

when the profile is added to the database.

When the profiles contain the correct information, press <***OK***> in the *Update MLST alleles/profiles* dialog box to add all new profiles in the list to the database.

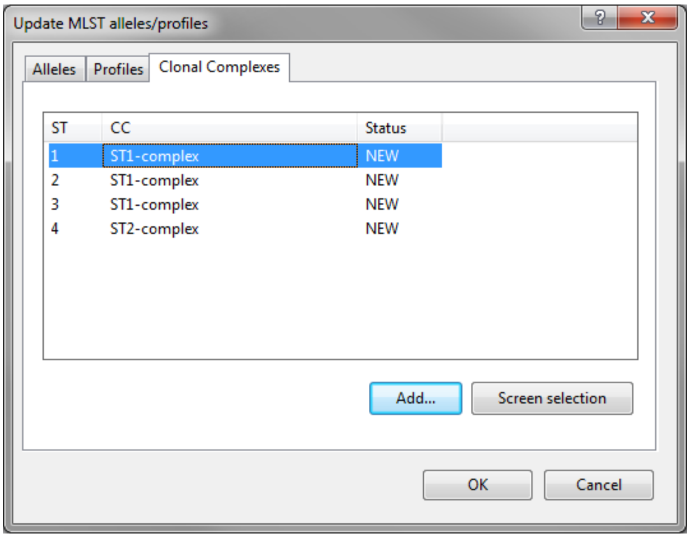
In a similar fashion as for alleles, profiles can be manually added, updated or deleted from the database (see 5.5.2).

5.5.4 Editing clonal complexes

In the *Clonal Complexes tab*, clonal complex information can be added to the database.

When pressing the <***Screen selection***> button, the software compares the clonal complex and sequence type combinations of the current entry selection to the information that is already present in the custom database. All newly detected clonal complex and sequence type combinations are added to the list in the *Clonal Complexes tab*.

The ***ST*** column displays the name of the sequence type, the ***Clonal complex*** column holds the clonal complex information, and the last column displays the ***Status*** information.

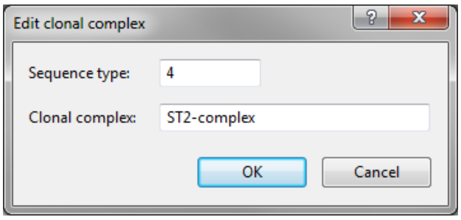


ST	CC	Status
1	ST1-complex	NEW
2	ST1-complex	NEW
3	ST1-complex	NEW
4	ST2-complex	NEW

Figure 5.11: Clonal complex information in the custom MLST database.

Double-click on a clonal complex and sequence type combination in the list if you want to edit the information. This pops up the *Edit clonal complex* dialog box.

A new clonal complex can also be added manually by pressing the <***Add***> button. This calls the *Edit clonal complex* dialog box (see Figure 5.12).



Sequence type: 4
Clonal complex: ST2-complex

Figure 5.12: Add clonal complex to the custom MLST database.

Pressing <***OK***> adds the new sequence type and clonal complex combination to the database.

When the clonal complexes contain the correct information, press <***OK***> to add all new clonal complexes in the list to the database.

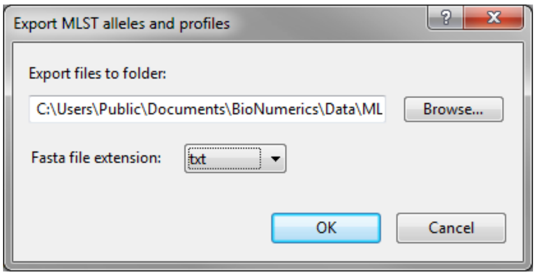
In a similar fashion as for alleles and profiles, clonal complexes can be manually added, updated or deleted from the database (see 5.5.2).

5.6 Exporting alleles and profiles

Locally stored allele and profile information can be exported as FASTA-formatted text files from the database with the command ***MLST*** > ***Export alleles and profiles***. This feature can be particularly useful when moving from an MLST setup with the allele and profile information stored in the database to a setup in which the information is made available through files, stored on a network drive or located on the internet.

Browse for an export folder and select a ***FASTA file extension*** from the drop down list.

Pressing <***OK***> exports the MLST allele and profile information. For each sequence type, BioNumerics creates a separate FASTA file, containing the sequences that are locally stored in the database. The allelic profiles, sequence types and clonal complex information are stored in the tab-delimited `Profiles.txt` file.


Export MLST alleles and profiles

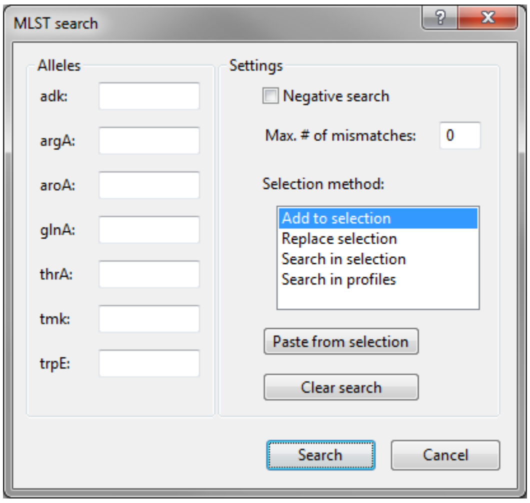
Export files to folder: C:\Users\Public\Documents\BioNumerics\Data\ML

Fasta file extension: txt

Figure 5.13: The *Export MLST alleles and profiles* dialog box: Export locally stored allele and profile information.

5.7 MLST search

The MLST search tool is designed for searching entries with certain allelic profiles, corresponding to combinations of character states in the **MLST** character data set. Selecting ***MLST*** > ***MLST search*** calls the *MLST search* dialog box (see Figure 5.14).


MLST search

Alleles

adk:

argA:

aroA:

glnA:

thrA:

tmk:

trpE:

Settings

☐ Negative search

Max. # of mismatches: 0

Selection method: Add to selection

Figure 5.14: The *MLST search* dialog box.

The *MLST search* dialog box lists all loci that are defined in the MLST scheme. The user can enter the selection criteria for each of the loci in the *Alleles panel*. The criteria for the different loci are combined with AND. For each locus, comma separated values can be specified, which are then combined with OR.

The selection criteria can be cleared with the <***Clear search***> button. In the right panel, a ***Maximum number of mismatches*** can be specified. With the ***Negative search*** option checked, all entries that do not match the specified criteria will be selected. One can specify to add the entries to the current selection (***Add to selection***), replace an existing selection of entries by the found entries (***Replace selection***), or to search within an existing selection of entries (***Search in selection***). With ***Search in profiles***, the profiles in the database are searched. Search criteria can be pasted from the character states of the currently selected entries using the <***Paste from selection***> button.

When pressing the <***Search***> button all entries in the database that match the specified criteria are selected and are brought to the top of the *Database entries* panel. In case ***Search in profiles*** was selected, a window will pop up, listing all sequence types that fulfilled the search criteria.

Chapter 6

MLST data analysis

Some useful features in the context of MLST data analysis will be highlighted in this chapter. More detailed information about the analysis possibilities in the software can be found in the BioNumerics manual.

6.1 Selections in BioNumerics

1.1 Select a single entry in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

Selected entries are marked by a checked ballot box (☑) and can be unselected in the same way.

1.2 In order to select a group of entries, hold the **Shift**-key and click on another entry.

A group of entries can be unselected the same way.

1.3 All entries can be selected at once with ***Edit*** > ***Select all*** (**Ctrl+A**).

1.4 Clear all selected entries with ***Database*** > ***Entries*** > ***Unselect all entries (all levels)*** (, **F4**).

6.2 The Comparison window

2.1 Make a selection in the *Database entries* panel (see 6.1).

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

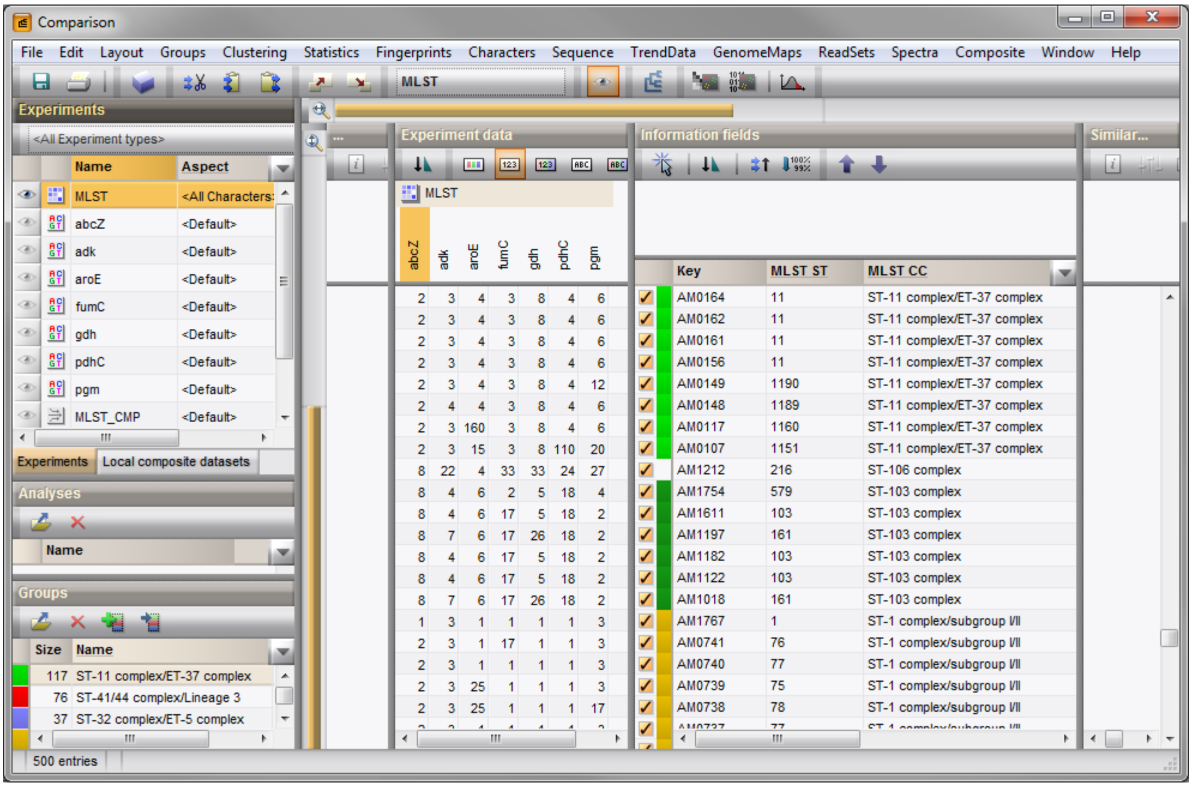
A *Comparison* window is created, with the selected database entries.

2.3 You can drag the vertical separator lines between the panels to the left or to the right, in order to divide the space among the panels optimally.

2.4 Click on the 👁 next to the experiment name **MLST** in the *Experiments* panel and select ***Characters*** > ***Show values*** (123) to display the allele numbers in the *Experiment data* panel (see Figure 6.1).

2.5 In the *Comparison* window, groups can be created based on a selection (***Groups*** > ***Create new group from selection*** (, **Ctrl+G**)) or based on the content of an information field: right-click in the header of a field (e.g. field 'MLST CC') and select ***Create groups from database field*** from the menu.

2.6 Click on the 👁 next to a sequence experiment name in the *Experiments* panel to view the sequences in the *Experiment data* panel.



Name	Aspect
MLST	<All Characters>
abcZ	<Default>
adk	<Default>
aroE	<Default>
fumC	<Default>
gdh	<Default>
pdhC	<Default>
pgm	<Default>
MLST_CMP	<Default>

Size	Name
117	ST-11 complex/ET-37 complex
76	ST-41/44 complex/Lineage 3
37	ST-32 complex/ET-5 complex

abcZ	adk	aroE	fumC	gdh	pdhC	pgm		Key	MLST ST	MLST CC
2	3	4	3	8	4	6	☑	AM0164	11	ST-11 complex/ET-37 complex
2	3	4	3	8	4	6	☑	AM0162	11	ST-11 complex/ET-37 complex
2	3	4	3	8	4	6	☑	AM0161	11	ST-11 complex/ET-37 complex
2	3	4	3	8	4	6	☑	AM0156	11	ST-11 complex/ET-37 complex
2	3	4	3	8	4	12	☑	AM0149	1190	ST-11 complex/ET-37 complex
2	4	4	3	8	4	6	☑	AM0148	1189	ST-11 complex/ET-37 complex
2	3	160	3	8	4	6	☑	AM0117	1160	ST-11 complex/ET-37 complex
2	3	15	3	8	110	20	☑	AM0107	1151	ST-11 complex/ET-37 complex
8	22	4	33	33	24	27	☑	AM1212	216	ST-106 complex
8	4	6	2	5	18	4	☑	AM1754	579	ST-103 complex
8	4	6	17	5	18	2	☑	AM1611	103	ST-103 complex
8	7	6	17	26	18	2	☑	AM1197	161	ST-103 complex
8	4	6	17	5	18	2	☑	AM1182	103	ST-103 complex
8	4	6	17	5	18	2	☑	AM1122	103	ST-103 complex
8	7	6	17	26	18	2	☑	AM1018	161	ST-103 complex
1	3	1	1	1	1	3	☑	AM1767	1	ST-1 complex/subgroup I/II
2	3	1	17	1	1	3	☑	AM0741	76	ST-1 complex/subgroup I/II
2	3	1	1	1	1	3	☑	AM0740	77	ST-1 complex/subgroup I/II
2	3	25	1	1	1	3	☑	AM0739	75	ST-1 complex/subgroup I/II
2	3	25	1	1	1	17	☑	AM0738	78	ST-1 complex/subgroup I/II

Figure 6.1: The *Comparison* window: groups have been defined based on the ***MLST CC*** information field.

2.7 Click on the 👁 next to the experiment name **MLST_CMP** in the *Experiments* panel to view the mutation list of the concatenated MLST sequences in the *Experiment data* panel.

6.3 Cluster analysis

3.1 In the *Experiments* panel of the *Comparison* window, make sure the **MLST** experiment is selected.

3.2 Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** to display the *Similarity coefficient* wizard page.

Due to the arbitrariness of the allele numbers, the only suitable similarity coefficients for clustering MLST data are categorical coefficients. A categorical coefficient compares the allele numbers to see if they are the same or different but does not quantify the difference.

3.3 Select e.g. the ***Categorical (values)*** coefficient from the list and press <***Next***>.

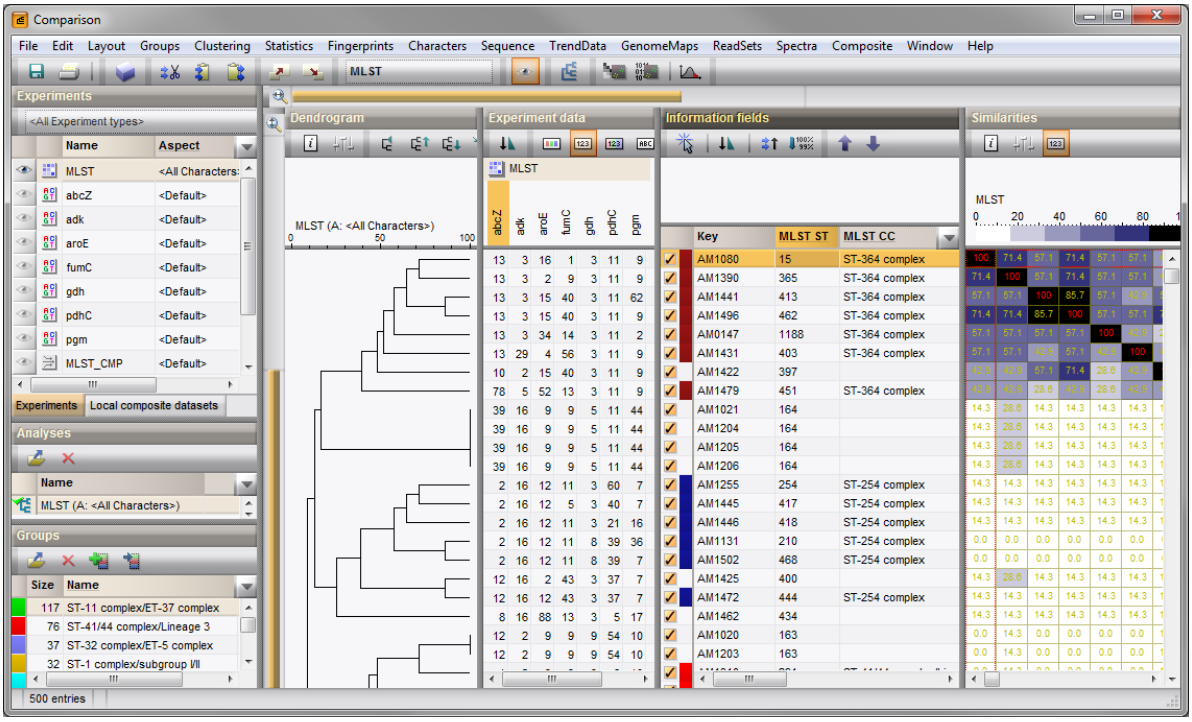
3.4 In the *Cluster analysis* wizard page, choose a clustering method (e.g. **UPGMA**) and press <***Finish***>.

When finished, the dendrogram and the similarity matrix are shown in the *Comparison* window (see Figure 6.2 for an example).

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

3.5 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


Comparison
File Edit Layout Groups Clustering Statistics Fingerprints Characters Sequence TrendData GenomeMaps ReadSets Spectra Composite Window Help
MLST
Experiments
<All Experiment types>
Experiments Local composite datasets
Analyses
Groups
Dendrogram
MLST (A: <All Characters>)
0 50 100
Experiment data
MLST
Information fields
Similarities
MLST
0 20 40 60 80

Name	Aspect
MLST	<All Characters>
abcZ	<Default>
adk	<Default>
aroE	<Default>
fumC	<Default>
gdh	<Default>
pdhC	<Default>
pgm	<Default>
MLST_CMP	<Default>

Name
MLST (A: <All Characters>)

Size	Name
117	ST-11 complex/ET-37 complex
76	ST-41/44 complex/Lineage 3
37	ST-32 complex/ET-5 complex
32	ST-1 complex/subgroup I/II

500 entries

abcZ	adk	aroE	fumC	gdh	pdhC	pgm		Key	MLST ST	MLST CC
13	3	16	1	3	11	9	☑	AM1080	15	ST-364 complex
13	3	2	9	3	11	9	☑	AM1390	365	ST-364 complex
13	3	15	40	3	11	62	☑	AM1441	413	ST-364 complex
13	3	15	40	3	11	9	☑	AM1496	462	ST-364 complex
13	3	34	14	3	11	2	☑	AM0147	1188	ST-364 complex
13	29	4	56	3	11	9	☑	AM1431	403	ST-364 complex
10	2	15	40	3	11	9	☑	AM1422	397	
78	5	52	13	3	11	9	☑	AM1479	451	ST-364 complex
39	16	9	9	5	11	44	☑	AM1021	164	
39	16	9	9	5	11	44	☑	AM1204	164	
39	16	9	9	5	11	44	☑	AM1205	164	
39	16	9	9	5	11	44	☑	AM1206	164	
2	16	12	11	3	60	7	☑	AM1255	254	ST-254 complex
2	16	12	5	3	40	7	☑	AM1445	417	ST-254 complex
2	16	12	11	3	21	16	☑	AM1446	418	ST-254 complex
2	16	12	11	8	39	36	☑	AM1131	210	ST-254 complex
2	16	12	11	8	39	7	☑	AM1502	468	ST-254 complex
12	16	2	43	3	37	7	☑	AM1425	400	
12	16	12	43	3	37	7	☑	AM1472	444	ST-254 complex
8	16	88	13	3	5	17	☑	AM1462	434	
12	2	9	9	9	54	10	☑	AM1020	163	
12	2	9	9	9	54	10	☑	AM1203	163	

Figure 6.2: The *Comparison* window with a UPGMA tree displayed.

the similarity matrix, and will calculate a standard minimum spanning tree with single and double locus variance priority rules.

3.6 Specify an analysis name (for example **MLST1**), make sure ***MLST(<All Characters>)*** is selected, select ***MST for categorical data***, and press <***Next***>.

The *Advanced cluster analysis* window pops up (see Figure 6.3 for an example).

The *Network panel* in the *Advanced cluster analysis* window 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 (if defined) are automatically shown as node colors.

3.7 To change the display settings of the network (node/branch labels, node/branch colors, etc.), press the button or choose ***Display*** > ***Display settings***.

3.8 Select ***File*** > ***Exit*** to close the *Advanced cluster analysis* window and select ***File*** > ***Save*** (, **Ctrl+S**) to save the comparison.

All calculations done on the data is stored along with the comparison. This includes a similarity matrix for the experiment type on which a dendrogram was last calculated, any dendrogram that has been created, any statistical analysis, comparison groups, etc..

3.9 Enter a name, e.g. "MyComp" and press <***OK***>.

3.10 Close the comparison with ***File*** > ***Exit***. The comparison **MyComp** is now listed in the *Comparisons* panel of the *Main* window.

To continue working on an existing comparison, it can be opened again:

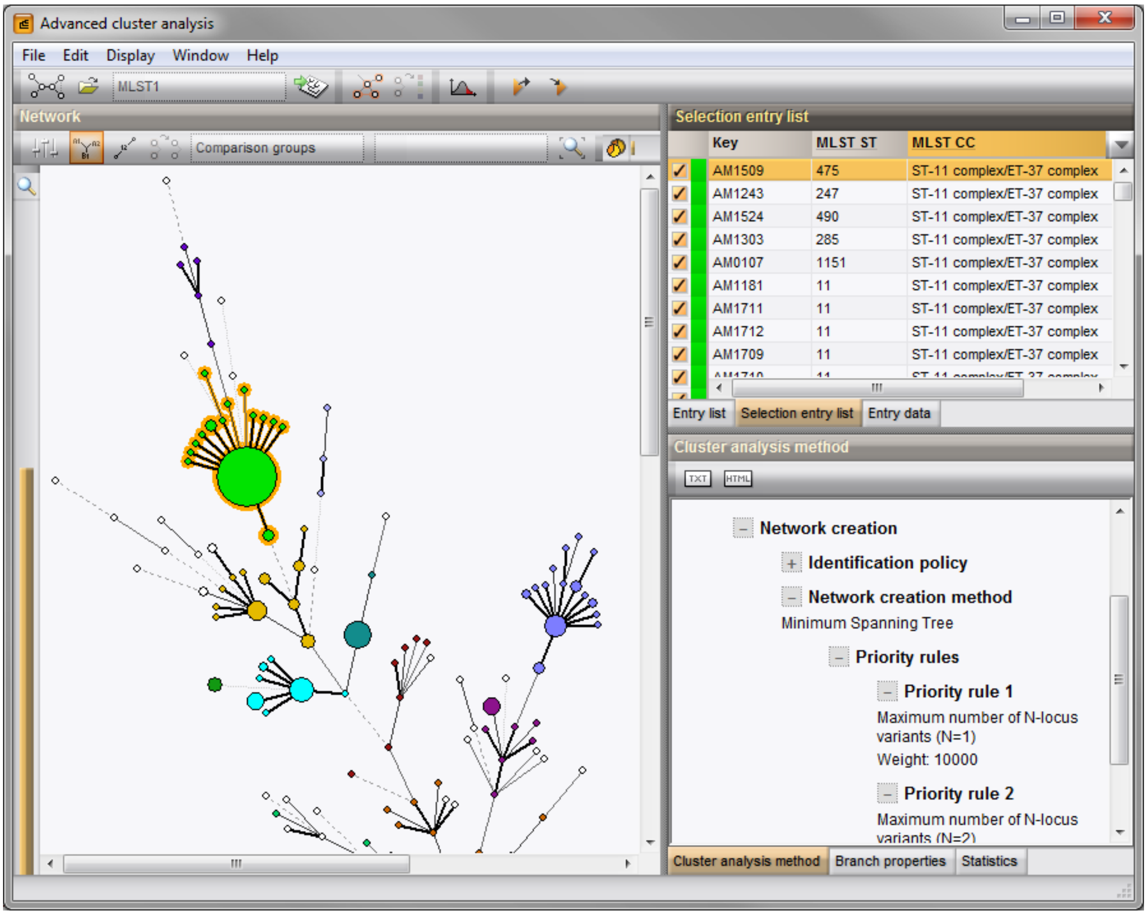
3.11 Highlight the comparison in the *Comparisons* panel and select ***Edit*** > ***Open highlighted object...*** (, **Enter**). Alternatively, just double-click on the comparison name.

A BIOMÉRIEUX COMPANY

Please contact us for any
additional information you might
require, we will gladly help you!

Headquarters
Keistraat 120 • 9830 Sint-Martens-Latem • Belgium
+32 922 22 100 info@applied-maths.com

USA and Canada
11940 Jollyville Rd., Suite 115N • Austin, TX 78750 USA
+1 512 482 9700 info-us@applied-maths.com



	Key	MLST ST	MLST CC
✓	AM1509	475	ST-11 complex/ET-37 complex
✓	AM1243	247	ST-11 complex/ET-37 complex
✓	AM1524	490	ST-11 complex/ET-37 complex
✓	AM1303	285	ST-11 complex/ET-37 complex
✓	AM0107	1151	ST-11 complex/ET-37 complex
✓	AM1181	11	ST-11 complex/ET-37 complex
✓	AM1711	11	ST-11 complex/ET-37 complex
✓	AM1712	11	ST-11 complex/ET-37 complex
✓	AM1709	11	ST-11 complex/ET-37 complex

- Network creation
  - Identification policy
  - Network creation method
    Minimum Spanning Tree
    - Priority rules
      - Priority rule 1
        Maximum number of N-locus variants (N=1)
        Weight: 10000
      - Priority rule 2
        Maximum number of N-locus variants (N=2)

Figure 6.3: The *Advanced cluster analysis* window.