BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

Spa typing plugin

Contents

1 Starting and setting up BIONUMERICS **5**
- 1.1 Introduction . . . 5
- 1.2 Startup program . . . 5
- 1.3 Creating a new database . . . 5
- 1.4 Installing the Spa typing plugin . . . 7

2 Getting started **11**
- 2.1 Browsing Spa repeats or types . . . 11
- 2.2 Updating Spa types and Spa repeats . . . 13

3 Importing and assembling trace files **15**
- 3.1 Importing and assembling trace files in batch . . . 15
- 3.2 Reports . . . 18

4 Checking assemblies in Assembler **21**
- 4.1 Introduction . . . 21
- 4.2 Showing Spa repeats on the consensus . . . 22
- 4.3 Showing the repeat succession plot . . . 24
- 4.4 Changing the status of error (and warning) messages . . . 26
  - 4.4.1 Principles . . . 26
  - 4.4.2 Option1: Changing the status in Assembler . . . 27
  - 4.4.3 Option2: Changing the status in the Detailed report window . . . 27

5 Spa-Typing in BIONUMERICS **29**
- 5.1 Selections in the main window . . . 29
- 5.2 Assigning Spa types . . . 29
  - 5.2.1 Principles . . . 29
  - 5.2.2 Step 1: The assembly is screened for repeats . . . 30
  - 5.2.3 Step 2: Repeat type (if available) is assigned to each selected entry . . . 31

6 Cluster analysis of Spa types **33**
- 6.1 Introduction . . . 33
- 6.2 The Comparison window . . . 33
- 6.3 Creating a cost matrix . . . 34
- 6.4 Cluster analysis settings . . . 36
- 6.5 Minimum spanning tree . . . 37
- 6.6 Cluster analysis sensu stricto . . . 38

7 Matching Spa types **41**
- 7.1 Selections in the main window . . . 41
- 7.2 Matching Spa types . . . 41

8 Synchronizing with SpaServer **45**
- 8.1 SpaServer information fields . . . 45

8.2 SpaServer synchronization settings 47
8.2.1 Introduction 47
8.2.2 Add SpaServer users 47
8.2.3 Link BIONUMERICS information fields to SpaServer fields 49
8.2.4 Store SpaServer results in BIONUMERICS database fields 49
8.3 Synchronizing with SpaServer (batch mode) 50
8.4 Synchronizing with SpaServer (entry mode) 52

NOTES

SUPPORT BY APPLIED MATHS, A BIOMÉRIEUX COMPANY

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, a bioMérieux company, 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: BE-DAU-INFO@biomerieux.com
URL: https://www.bionumerics.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: US-DAU-INFO@biomerieux.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:

- Python 3.8 release from the Python Software Foundation, https://www.python.org/
- Xerces library for XML input and output from the Apache Software Foundation, https://xerces.apache.org/
- NCBI toolkit version 2.11.0, https://www.ncbi.nlm.nih.gov/BLAST/
- SRA Toolkit, https://ncbi.github.io/sra-tools/
- Boost c++ libraries, https://www.boost.org/
- Samtools for interacting with SAM / BAM files, https://www.htslib.org/download/
- 7-Zip (7za.exe), https://www.7-zip.org/
- Zlib library, https://zlib.net/
- Pigz for parallel gzip compression, https://zlib.net/pigz/
- Cairo 2D graphics library version 1.12.14, https://cairographics.org/
- Crypto++ library version 5.5.2, https://www.cryptopp.com/
- OpenSSL library, https://www.openssl.org/
- libSVM library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SQLite version 3.7.17, https://www.sqlite.org/
- pymzML Python module version 2.4.7, https://github.com/pymzml/pymzML
- NumPy Python library version 1.19.1, https://www.numpy.org/
- BioPython Python library version 1.78, https://www.biopython.org/
- pyodbc Python module version 4.0.30, https://pypi.org/project/pyodbc/
- jinja2 Python library version 2.11.2, https://pypi.org/project/Jinja2/
- MarkupSafe Python library version 1.1.1, https://pypi.org/project/MarkupSafe/
- regex Python library version 2.5.91, https://pypi.org/project/regex/
- Chromium Embedded Framework, https://bitbucket.org/chromiumembedded/cef/wiki/Home
- SPAdes genome assembler version 3.15.3, https://bioinf.spbau.ru/spades *
- SKESA version 2.3.0, https://github.com/ncbi/SKESA/releases
- Unicycler version 0.5.0, https://github.com/rrwick/Unicycler/releases *
- Velvet for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Bowtie2 version 2.2.5 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml)*
- SNAP version 2.0.0, https://www.microsoft.com/en-us/research/project/snap/
- RAxML version 8.2.11, https://github.com/stamatak/standard-RAxML/releases

- FastTree version 2.1.10, https://www.microbesonline.org/fasttree/
- CFSAN SNP pipeline version 2.2.0, https://github.com/CFSAN-Biostatistics/snp-pipeline *
- Prokka version 1.14.5, https://github.com/tseemann/prokka *
- sourmash version 4.1.0, https://github.com/dib-lab/sourmash **
- SeqSero2 for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Fastp version 0.22.0, https://github.com/OpenGene/fastp

*: On Calculation Engine only **: See license conditions below

Sourmash license conditions:

Chapter 1

Starting and setting up BIONUMERICS

1.1 Introduction

This guide is designed as a tutorial for the *Spa typing plugin*. This plugin offers extra functionality to BIONUMERICS to do Spa typing for *Staphylococcus aureus*. Sequences in the database can be screened for known spa repeats and types downloaded from the SpaServer, data can be submitted to the SpaServer via a synchronization process, and entries can be clustered based on the spa types.

The features of the plugin will be illustrated using data available on the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "SPA typing data files"). The *Spa typing plugin* is supported in the **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE**.

1.2 Startup program

Make sure the latest version of BIONUMERICS is installed (https://www.bionumerics.com/download/software). The installation manual can be downloaded from https://www.bionumerics.com/download/manuals.

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.

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

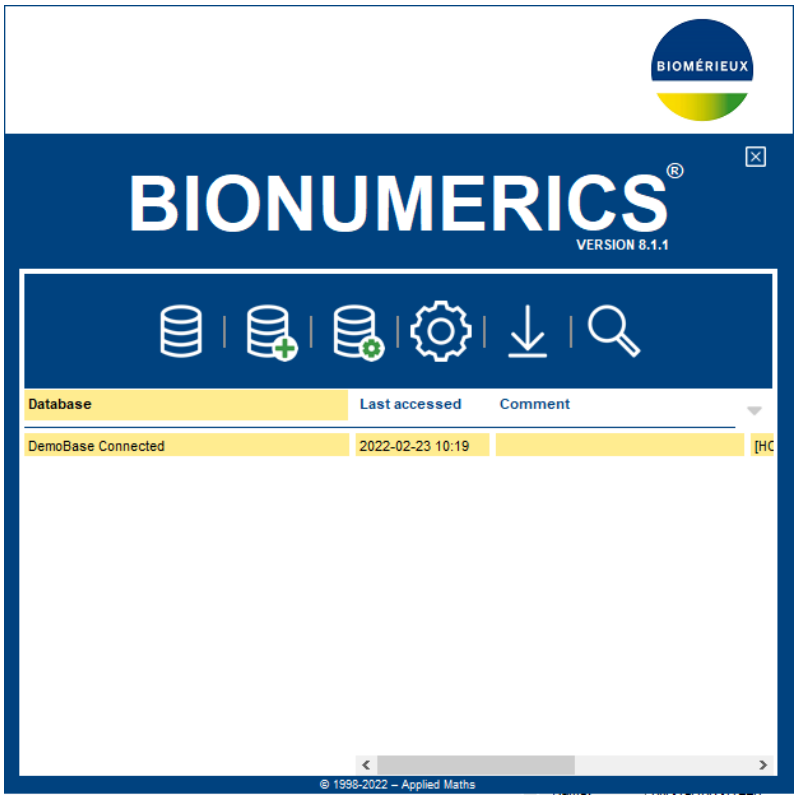


Figure 1.1: The *BIONUMERICS Startup* window.

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

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

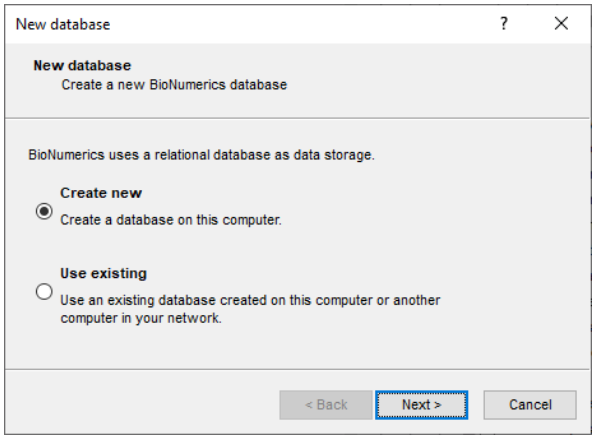


Figure 1.2: The *New database* wizard page.

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

3.4 Leave the default option selected and press <***Finish***> to complete the setup of the new database.

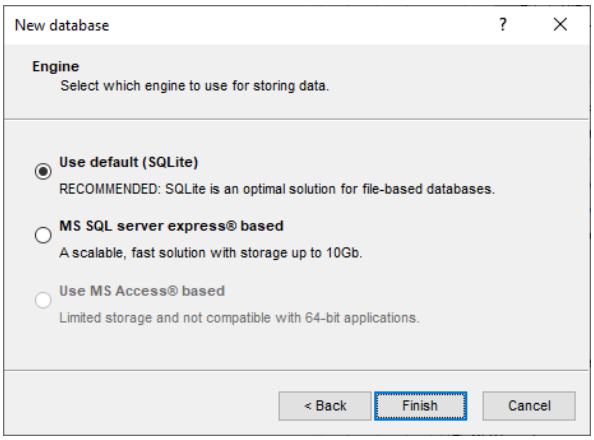


Figure 1.3: The *Engine* wizard page.

1.4 Installing the Spa typing plugin

The *Plugins and Scripts* dialog box can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ) (see Figure 1.4).

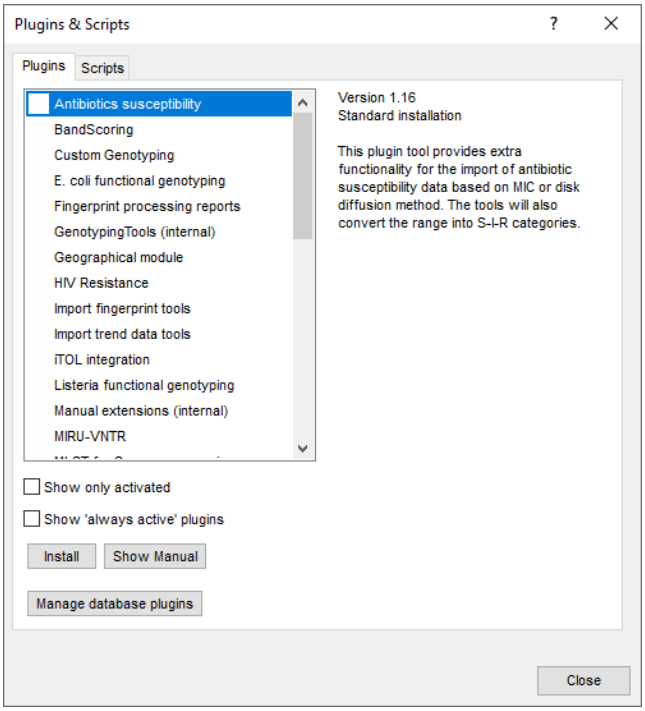


Figure 1.4: The *Plugins and Scripts* dialog box.

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 <***Install***> button. The software will ask for confirmation before installation. Some plugins are only supported in specific BIONUMERICS configurations. If the plugin is not supported by your BIONUMERICS configuration, it 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 <***Uninstall***> button.

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

4.1 Select the *Spa typing plugin* from the list and press the <***Install***> button.

4.2 The program will ask to confirm the installation of the plugin. Press <***OK***> twice to confirm the installation.

The *Spa typing settings* dialog box pops up (see Figure 1.5).

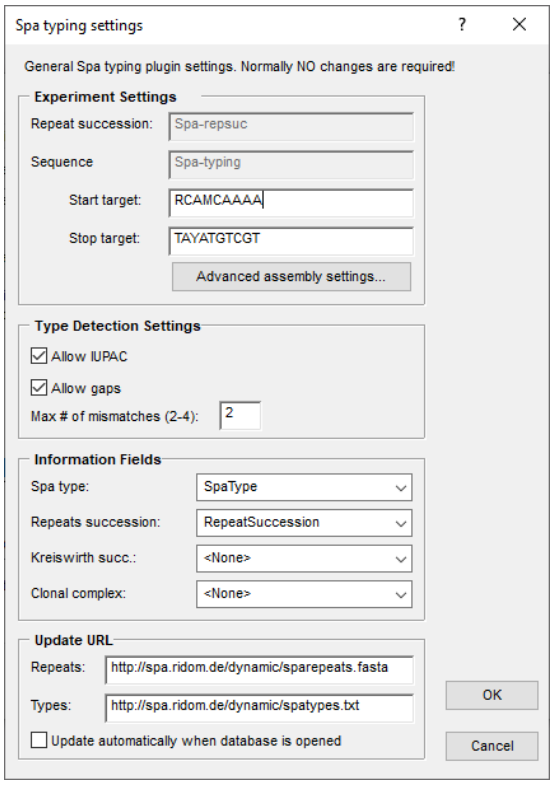


Figure 1.5: The *Spa typing settings* dialog box.

Experiment Settings:

- The sequence type **Spa-typing** is automatically created upon installation of the *Spa typing plugin*, and will be used for the storage of the imported sequences (***Sequence***).
- The Spa repeat successions are stored in the character type experiment **Spa-repsuc** (***Repeat succession***).
- The start and stop trim patterns are automatically filled out in the ***Start*** and ***Stop target*** boxes respectively, but can be changed if desired.

If you want to submit sequences to the online SpaServer (see 8), the sequences must contain the following signatures: 5' signature: `RCAMCAAAA`, 3' signature: `TAYATGTCGT`.

When pressing the <***Advanced Assembly settings***> button, the *Assembly settings* dialog box pops up (see Figure 1.6).

The Assembly settings are grouped per settings dialog box in Assembler: ***Quality Assignment***, ***Calculate Assembly***, and ***Consensus Determination***. For a detailed description of the Assembler program settings, see the reference manual.

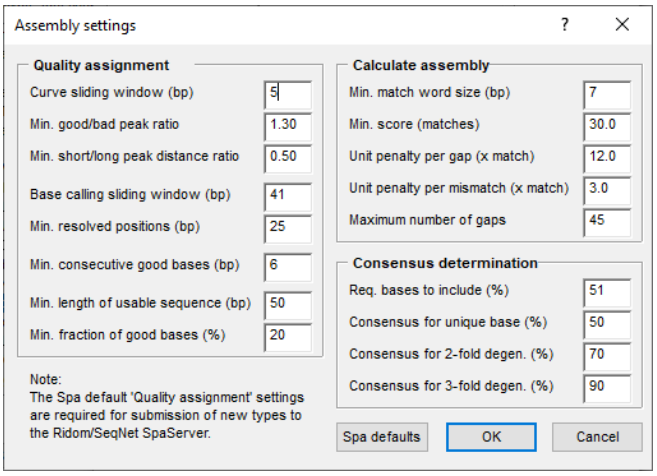


Figure 1.6: The *Assembly settings* dialog box for Spa Typing.

The Assembly settings can still be changed after installation of the plugin with ***Spa-Typing*** $>$ ***Settings***.

The default Spa ***Quality Assignment*** settings are required for submission of new types to the SpaServer (see 8). Pressing the $<$***Spa defaults***$>$ button will reset all settings in to their defaults.

Type Detection Settings:

- ***Allow IUPAC***: When this option is enabled, the tool will consider the different possibilities for the ambiguous positions for the repeat calling in Assembler. (see 4.2). This option is enabled by default.
- ***Allow gaps***: When this option is checked, gaps are allowed when searching for possible repeats in the consensus sequence.
- ***Maximum number of mismatches***: In the *Spa typing plugin*, a visualization tool is available (see 4.3) with editing suggestions for the unknown repeat(s). With this option, you can specify the maximum number of mismatches you want to consider between the source sequence and the repeat sequence. The maximum value is 4. Entering a higher number will cause the value to be set to 4.

Information Fields:

In the *Information fields panel*, you can choose the names of the database information fields that will contain the ***Spa type***, the ***Repeat succession***, the ***Kreiswirth succession*** string, and ***Clonal complex*** information for the entries in the database (see Figure 1.5). You can choose the default suggested names, select an existing field, enter a new field name or set the box to ”None”.

The storage of a repeat succession in an information field is used for illustration purposes only. Long repeat successions may be truncated due to size limitations of the information field. The repeat information stored in the associated character type will be used when using the matching and clustering tools.

If you want to change the name of one of the information fields selected in the *Information fields panel*, you need to rename the information fields in the database **and** in the *Information fields panel* in order to run the plugin tool properly.

A new information field can not start with a space.

Update URL:

- The URL for the update of the ***Repeats*** and the ***Types*** can be changed in the *Update URL panel*.
- When the option ***Update automatically when database is opened*** is checked, the software will automatically update all online repeat and type information each time the database is opened.

4.3 Leave all settings unaltered and press <***OK***>.

The software updates the types and repeats from the Ridom/Seqnet SpaServer.

4.4 When the *Spa typing plugin* is successfully installed, a confirmation message pops up. Press <***OK***> twice.

4.5 Press <***Proceed***> (or <***Close***>) to close the *Plugins and Scripts* dialog box and to continue to the *Main* window.

4.6 Close and reopen the database to activate the features of the *Spa typing plugin*.

The *Spa typing plugin* installs itself in a menu of the BIONUMERICS software. In the *Main* window, some initialization has been done to the database with the installation of the *Spa Typing plugin* (see Figure 1.7).

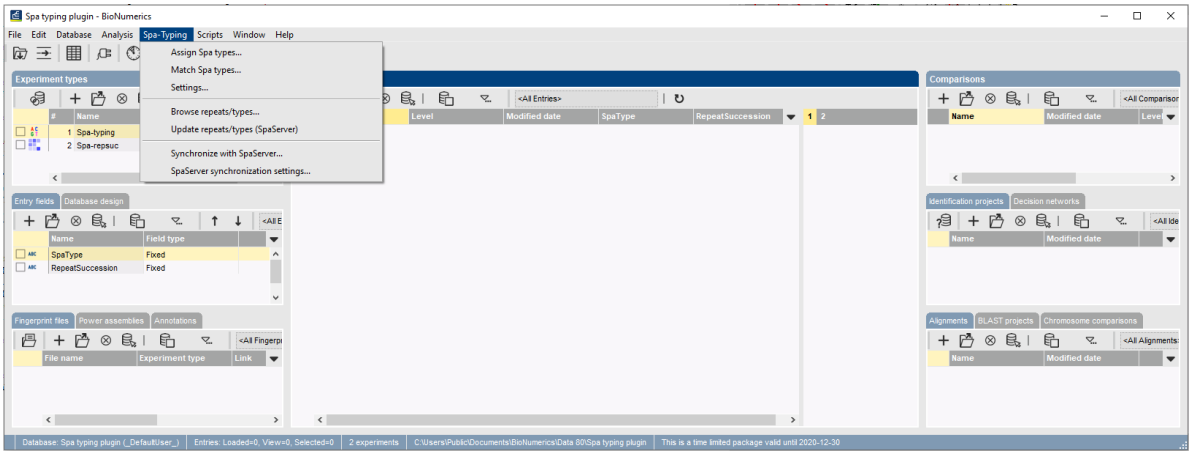


Figure 1.7: The *Main* window after installation of the *Spa Typing plugin*.

- The information fields specified in the *Information fields panel* of the *Spa typing settings* dialog box are present in the *Database entries* panel.
- The *Experiment types* panel lists the experiment types. BIONUMERICS has automatically created a sequence type called **Spa-typing** and a character type called **Spa-repsuc** upon installation of the plugin.

4.7 To call the *Spa typing settings* dialog box from the *Main* window, select ***Spa-Typing*** > ***Settings***.

4.8 Press the <***Advanced Assembly settings***> button to call the *Assembly settings* dialog box for the **Spa-typing** sequence experiment.

Chapter 2

Getting started

2.1 Browsing Spa repeats or types

The lists of Spa repeats and Spa types, downloaded from the SpaServer, can be queried by the user.

1.1 Select ***Spa-Typing*** > ***Browse repeats/types*** in the *Main* window.

This action calls the *Browse types/repeats* dialog box (see Figure 2.1).

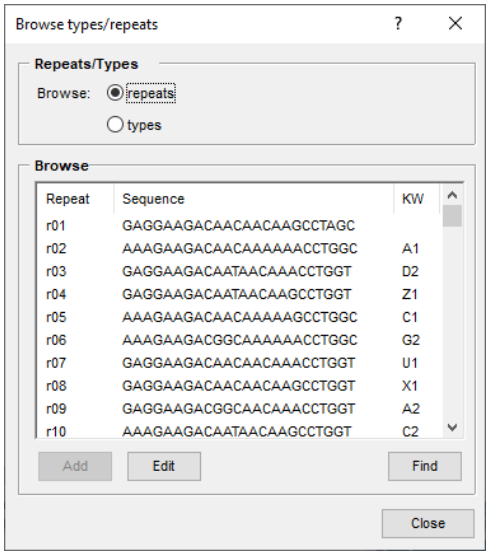


Figure 2.1: The *Browse types/repeats* dialog box.

In the *Repeats/Types panel* specify which list you want to browse: the ***repeats*** list or ***types*** list.

In the *Browse panel*, all repeats/types are listed that were downloaded from the online SpaServer.

Use the scroll bar to browse through the repeats/types.

Select repeats/types in the *Repeats/Types panel* and press the <***Find***> button to look for a repeat/type.

When looking for a repeat, enter the sequence in the *Find type* dialog box and press the <***Find***> button. If the repeat is present in the list of repeats, the ***Repeat ID*** of the sequence is displayed.

The ***Kreiswirth*** information box is only shown in the *Find type* dialog box if its corresponding information field is present in the database.

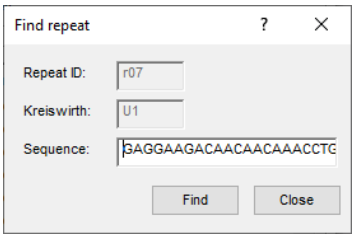


Figure 2.2: The *Find type* dialog box: Find repeat.

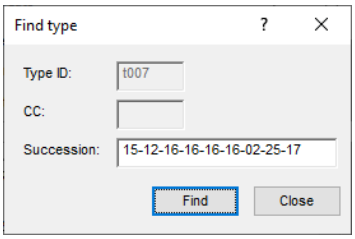


Figure 2.3: The *Find type* dialog box: Find type.

When looking for a type, enter the succession string in the *Find type* dialog box and press the <***Find***> button. If the succession string is present in the list of types, the ***Type ID*** of the succession string is displayed.

The ***Clonal complex*** information box is shown in the *Find type* dialog box if its corresponding information field is present in the database.

With the <***Add***> button, a new type can be added to the list of existing repeats/types.

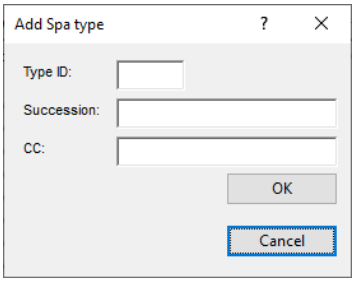


Figure 2.4: The *Add Spa type* dialog box.

Enter a ***Succession*** string, a ***Type ID*** and optional a clonal complex notation in the *Add Spa type* dialog box.

When pressing the <***OK***> button the type is added to the list of types.

When updating the spa repeat and spa type information from the SpaServer, types that were added manually to the database might be overwritten if the ***Type ID*** entered in the *Add Spa type* dialog box corresponds to a Type ID of a new online spa type.

Select a repeat/type in the *Browse panel* and press the <***Edit***> button to add/edit the Kreiswirth or clonal complex information.

If an information field name for the ***Kreiswirth*** information is specified in the *Information fields panel* of the *Spa typing settings* dialog box, Kreiswirth information can be linked in the *Edit Spa repeat* dialog box to the spa repeats that are present in the database.

If an information field name for the ***Clonal complex*** information is specified in the *Information*

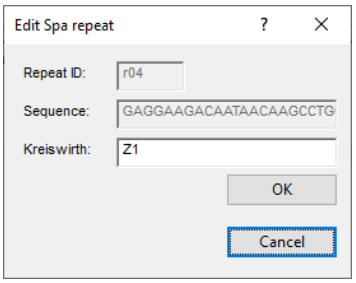


Figure 2.5: The *Edit Spa repeat* dialog box.

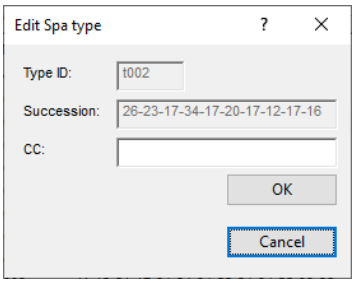


Figure 2.6: The *Edit Spa type* dialog box.

fields panel of the *Spa typing settings dialog box*, clonal complex information can be linked in the *Edit Spa type* dialog box to the spa types that are present in the database.

Clonal complex information that is stored outside BIONUMERICS (for example in an Excel or text file) can be imported into BIONUMERICS and linked to the spa types with a special script. Please contact Applied Maths to obtain this script.

It is also possible to view all repeats and types stored in the database with an *object query*.

1.2 In the *Main* window, select ***Database*** > ***Object queries...*** (▦) and select "<Create new>" from the drop-down menu that appears. Press <***OK***>.

1.3 As ***Object to report***, select "Spa repeats" or "Spa types" and press <***OK***> (see Figure 2.7).

For more information on object queries, we refer to the reference manual.

2.2 Updating Spa types and Spa repeats

When installing the *Spa typing plugin*, the known repeats and types are downloaded from the SpaServer.

2.1 The list of known repeats and types can be updated with the command ***Spa-Typing*** > ***Update repeats/types (SpaServer)***.

You can perform analyses without being connected to the internet, but you will be unable to update the list of known repeats and types.

When the option ***Update automatically when database is opened*** is checked in the *Spa typing settings* dialog box, BIONUMERICS automatically updates all online repeat and type information each time the database is opened.

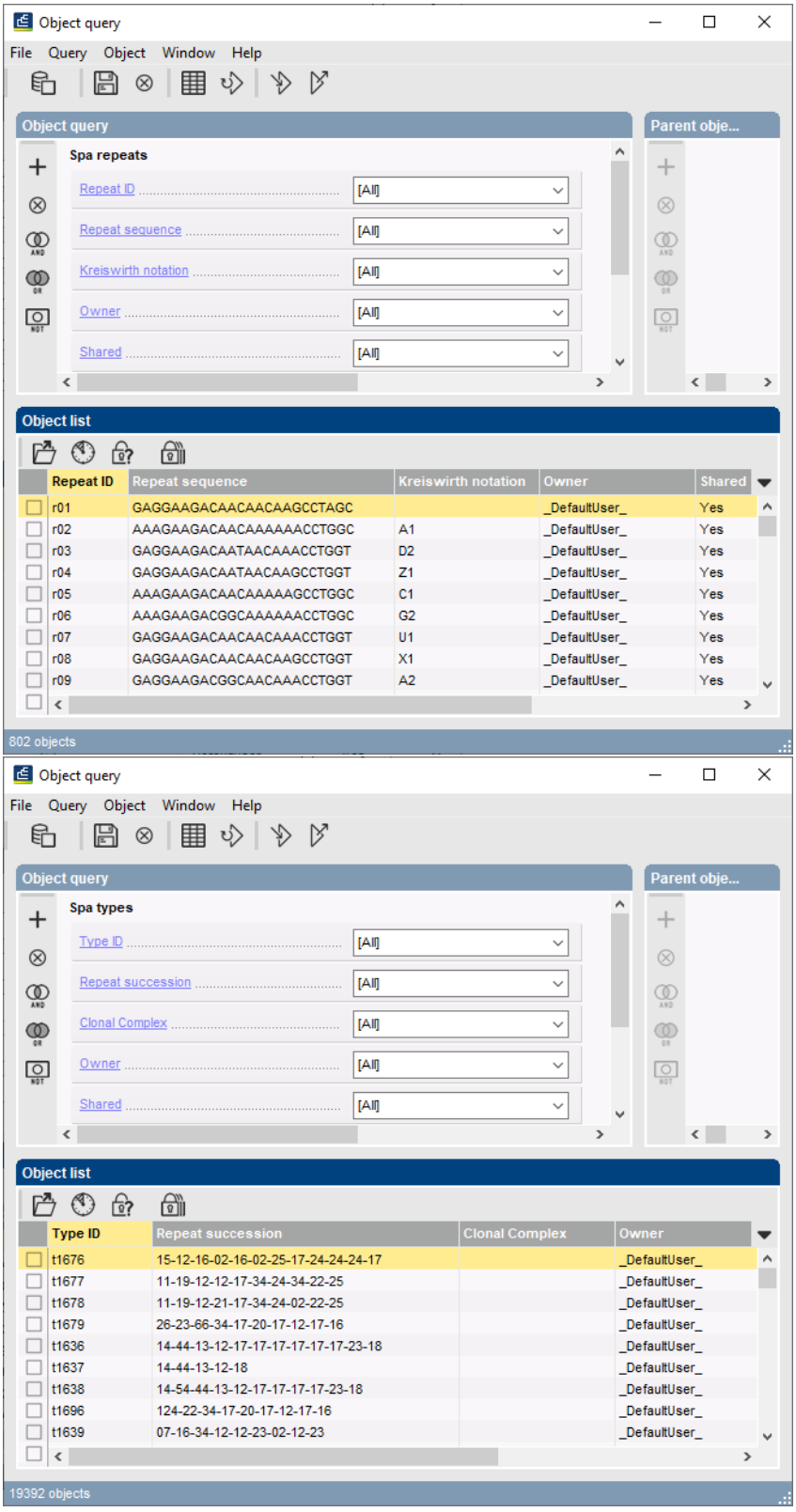


Figure 2.7: Object queries: Spa Repeats and Types.

Chapter 3

Importing and assembling trace files

3.1 Importing and assembling trace files in batch

A set of sequences run on an Applied Biosystems Genetic Analyzer can be downloaded from the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "SPA typing data files") and are used in this guide to explain the work flow of the *Spa-Typing plugin*.

1.1 Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import data* wizard.

1.2 Press the <***Browse***> button, navigate to the correct path, select all sequence trace files and press <***Open***>.

1.3 With the ***Import and assemble trace files*** option highlighted, press <***Finish***>.

The dialog is updated (see Figure 3.1).

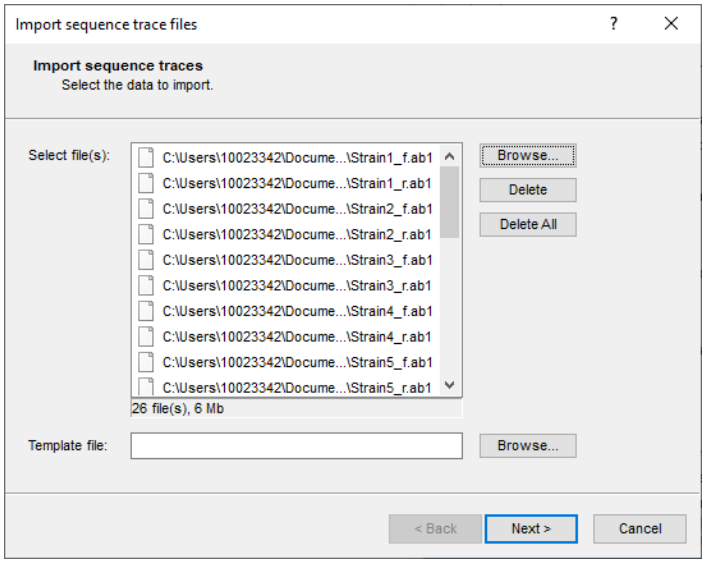


Figure 3.1: Select trace files.

1.4 Press <***Next***> to go the next step.

The way the information should be imported in the database can be specified with an import template. In the example data set, the ***Key*** is provided in the trace file name.

1.5 Make sure the ***Example import 1*** template is selected and press the <***Preview***> button.

The ***Example import 1*** template will parse the ***Key*** from the file names.

1.6 Close the preview.

1.7 Make sure the ***Example import 1*** template is selected, and select the ***Spa-typing*** from the ***Experiment type*** list (see Figure 3.2).

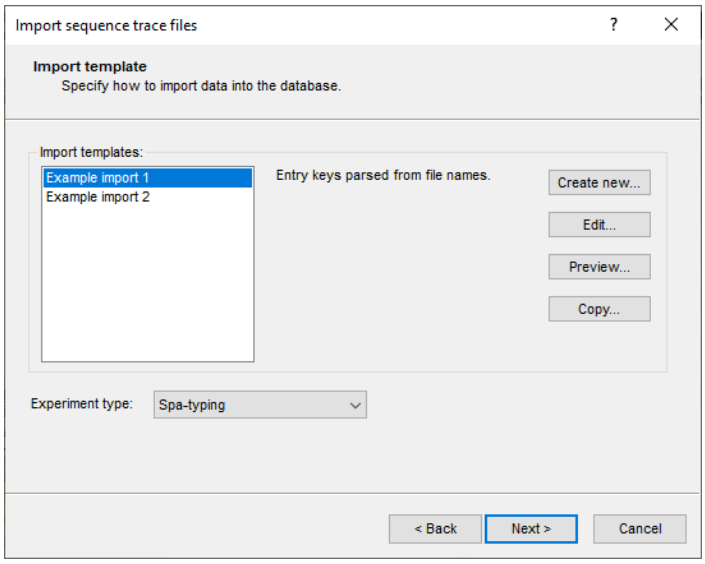


Figure 3.2: Import sequence trace files.

1.8 Press <***Next***>.

1.9 Press <***Next***> once more to confirm the creation of 13 new entries (see Figure 3.3).

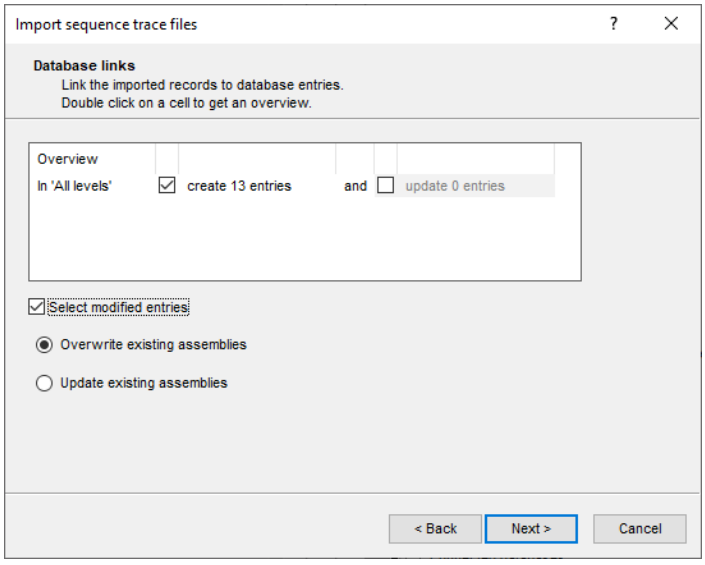


Figure 3.3: Database links.

The *Processing* dialog box opens (see Figure 3.4).

In the *Reports panel*, the ***Maximum# of unresolved bases reported*** can be specified (default value 20). Likewise, the ***Maximum # of align inconsistencies reported*** can be entered (default value 20). Align inconsistencies are positions where the consensus is resolved, but where one or

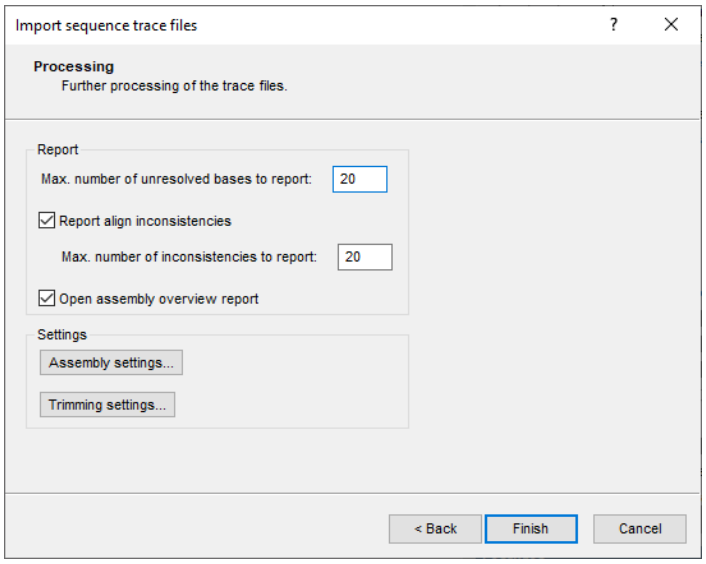


Figure 3.4: The *Processing* dialog box.

more sequences are different from the consensus.

1.10 Press <***Trimming settings***> to pop up the *Assembly trimming settings* dialog box (see Figure 3.5).

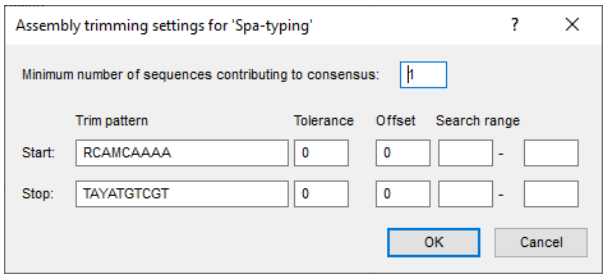


Figure 3.5: The *Assembly trimming settings* dialog box.

Following settings can be specified:

- ***Minimum # of sequences*** specifies the minimum number of trace sequences that should contribute to the subsequence on the consensus that matches the trimming targets. For example, if “2” is entered, a trimming target will only be set if the matching region on the consensus is *fully* defined by at least 2 sequences.
- For both the ***Start position*** and ***Stop position***, a ***Trim pattern*** is displayed. The use of IUPAC code for ambiguous positions is supported. The ***Tolerance*** defines the number of mismatches allowed for a sequence to be recognized as a trim pattern. With the ***Offset***, one can specify that the consensus is trimmed at a certain offset from the start and end trimming target positions. If no offset is specified (zero), the trimming targets are included in the trimmed consensus. With the ***Search range*** one can restrict the search to certain regions on the consensus, e.g. to prevent incidental matches inside the targeted consensus sequence.

The entered trim patterns will be searched on the consensus sequence in both directions, i.e. on the consensus as it appears as well as on its complementary strand. In case the trim patterns

match the complementary strand of the consensus, it will be automatically invert-complemented. If the ***Trim pattern*** text boxes are left empty, no preference sense is available.

The trimming patterns specified in the *Spa typing settings* dialog box (see Figure 1.5) are shown in the ***Start pattern*** and ***Stop pattern*** text boxes.

1.11 Leave the predefined settings unaltered and press <***OK***> to close the trimming dialog box.

1.12 Press the <***Assembly settings***> button to call the *Assembly settings* dialog box (see Figure 3.6).

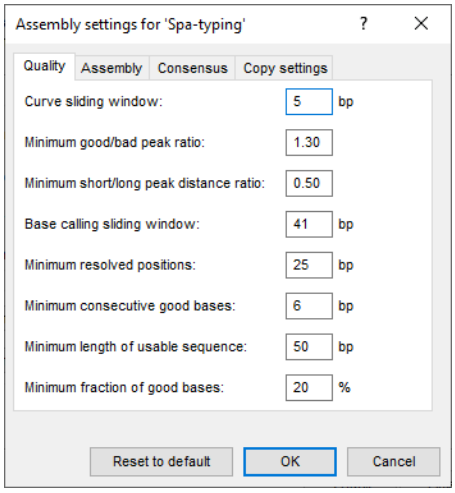


Figure 3.6: The *Assembly settings* dialog box.

The Assembly settings are grouped in tabs per settings dialog box in *Assembler*: ***Quality*** assignment, ***Assembly*** and ***Consensus*** determination. For a detailed description of the Assembler program settings, we refer to the reference manual. In the last tab the Assembly settings can be copied from or to another sequence type experiment.

The default Spa ***Quality*** assignment settings are required for submission of new types to the SpaServer (see 8).

1.13 For this exercise, do not change the settings and press <***OK***>.

1.14 Make sure the option ***Open assembly overview report*** is checked and press <***Finish***> to assemble the selected trace files from the example dataset into separate contig projects.

3.2 Reports

When the assemblies are processed, an interactive report window appears (see Figure 3.7). This window can also be displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

The *Overview* panel displays the entries (keys) as rows and the experiments as columns. Each cell, corresponding to a key/experiment pair, provides information about the current status of the contig project. This information can be:

- **N/A**: No such experiment exists with this key.
- **N/B**: An experiment with this key exists, but (a) the assembly was not created from this batch; or (b) no batch sequence assembly is present for this sequence.

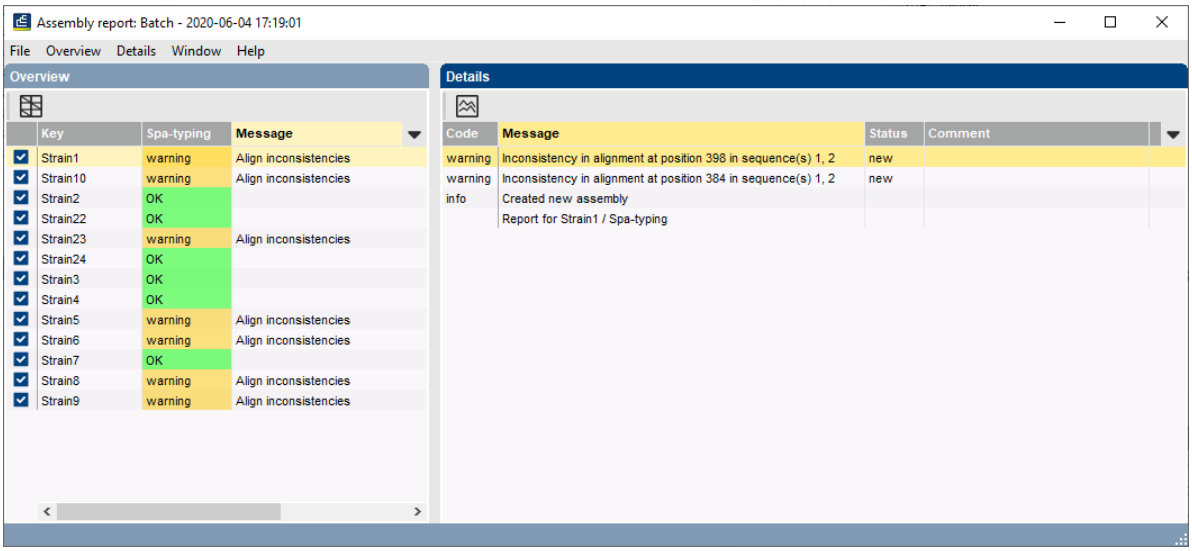


Figure 3.7: The *Batch sequence assembly report window* window.

- **OK** (green): A contig was assembled without any problems.
- **Warning** (orange): Align inconsistencies occurred that were resolved under the applied consensus determination settings.
- **Error** (red): At least one of several possible assembly errors occurred, e.g. a trace sequence did not meet the quality criteria, more than one contig was created, the trimming positions were not found or unresolved bases are present in the consensus.
- **Solved** (green): A warning or error that was solved by the user.

2.1 Click a cell, e.g. ***Strain23/Spa-typing*** to update the *Details* panel on the right-hand side.

The *Details* panel is organized in message rows with four columns.

- The first column displays a message **Code**, which can be either ”info”, ”warning” or ”error”.
- The second column shows the actual **Message**. Double-clicking on this cell opens the *Contig assembly* window (if not already open), with the corresponding position highlighted.
- The third column displays the **Status** of the message, which can be ”new”, ”read” or ”solved”. The status can be changed by the user.
- The fourth column is a **Comment** field. A comment can be entered by the user.

Chapter 4

Checking assemblies in Assembler

4.1 Introduction

The *Contig assembly* window can be launched from the *Batch sequence assembly report window* window or from the *Main* window:

- Double-click on a message cell in the *Details* panel*Details* panel of the *Batch sequence assembly report window* window of an key/experiment combination to launch Assembler.

- As soon as an experiment is linked to a database entry, the *Experiment presence* panel shows a colored dot for the experiment of this entry. Click on the colored dot in the *Experiment presence* panel while holding the **Shift**-key to open the *Experiment card* window for an entry. In the *Experiment card* window, click on the button to launch Assembler.

1.1 Open the *Contig assembly* window for the entry with key **Strain23** by double-clicking on the first message in the *Details* panel of the *Batch sequence assembly report window* window.

The *Alignment* panel in the *Contig assembly* window shows the consensus sequence (upper line) and the individual trace sequences that contribute to the displayed consensus. The upper panel (*Alignment overview* panel) displays the aligned trace sequences. If the arrow points to the left, the program has invert-complemented the sequence to obtain the correct alignment. The upper left panel displays the selected consensus with its length and the number of sequences that are part of it.

1.2 Select the *Aligned traces* panel.

The bottom panel now displays the chromatogram files for both trace sequences (see Figure 4.1).

1.3 To obtain an optimal view of the curves, use the zoom sliders in the *Traces* panel or use the zoom buttons.

The parameter ***Req. bases to include*** in the *Assembly settings* dialog box is by default set to 51% (see Figure 1.6). This means that a gap in one sequence and a nucleotide in the other will insert a gap in the consensus sequence. If you take a closer look at the alignment inconsistencies of this assembly, two gaps are present in the forward sequence (at positions 160 and 218), resulting in two gaps in the consensus sequence. These positions will be further investigated in the next steps.

Figure 4.1: The *Aligned traces* panel.

4.2 Showing Spa repeats on the consensus

2.1 In the *Contig assembly* window, select ***Spa-Typing*** > ***Show repeats*** or use the shortcut **Shift+F5**.

Assembler screens the consensus sequence for repeats.

- Known repeats are shown in *green* (see Figure 4.2) and the name of the repeat is shown on top of the known repeat sequence.
- Bases in the repeat succession string that are not assigned to a known repeat are shown in *red*.
- The 5' and 3' signatures are displayed in *yellow*.

If the option ***Allow IUPAC code*** is checked in the *Spa typing settings* dialog box (see Figure 1.5) and *one of the bases* of a IUPAC code in the consensus results in a match with a known repeat, the repeat is shown in green and the name of the repeat is shown on top of the corresponding repeat sequence in the *Alignment* panel.

If the option ***Allow IUPAC code*** is checked in the *Spa typing settings* dialog box (see Figure 1.5) and *more than one* of the bases of a IUPAC code in the consensus results in a match with a known repeat, the *Multiple repeat successions* dialog box displays the different repeat succession options.

The repeat of the selected match is shown in *orange* (see Figure 4.4) and the name of the matched repeat is shown on top of the corresponding repeat sequence followed by a question mark (e.g. "r12?").

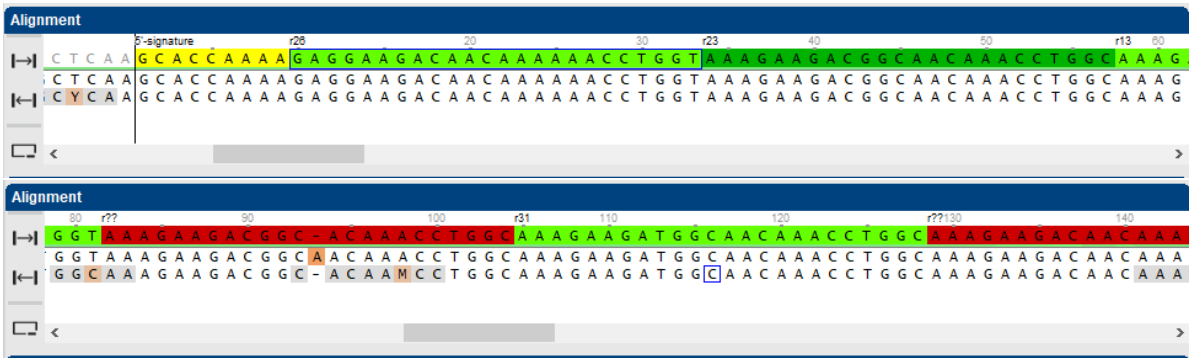


Figure 4.2: Showing the repeats on the consensus sequence.

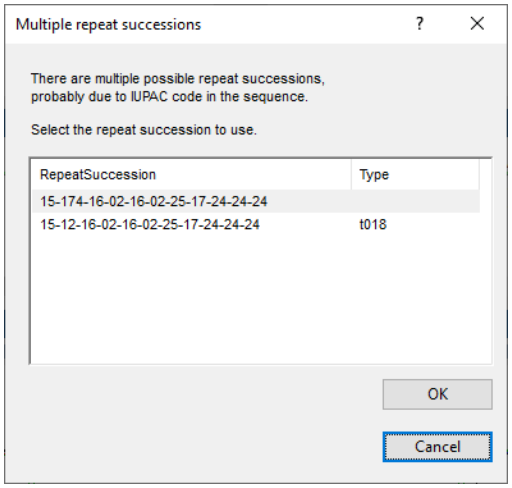


Figure 4.3: The *Multiple repeat successions* dialog box.

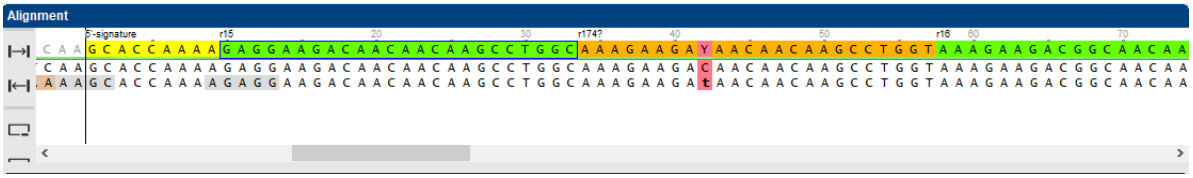


Figure 4.4: IUPAC code resulting in more than one known spa repeats.

The repeat succession string and the corresponding Spa type (if known) are displayed in the caption of the *Contig assembly* window (see Figure 4.5).

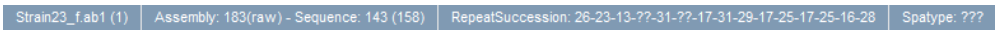


Figure 4.5: Repeat succession string.

When importing and assembling spa sequences, BIONUMERICS uses the parameters defined in the *Assembly settings* dialog box (see Figure 3.6).

2.2 Select ***File*** > ***Show report*** ( ) to view all parameters.

After import, these parameters can still be changed for each individual assembly.

1. Select the *Trimming* panel and select ***File*** > ***Quality assignment...*** ( ) to change the

quality assignment settings. This action can only be used if the alignment is removed.

2. Select the *Assembly* panel and choose ***Assembly*** > ***Assemble sequences...*** ( ) to change the assembly settings.

3. If you want to change the Consensus determination parameters, select the *Assembly* panel and select ***Assembly*** > ***Consensus determination...***.

Detailed information on each of these parameters can be found in the reference manual.

4.3 Showing the repeat succession plot

3.1 Select ***Spa-Typing*** > ***Show repeats plot*** or use the shortcut **Shift+F6**.

The repeats are displayed in the *Spa repeat plot window* (see Figure 4.6).

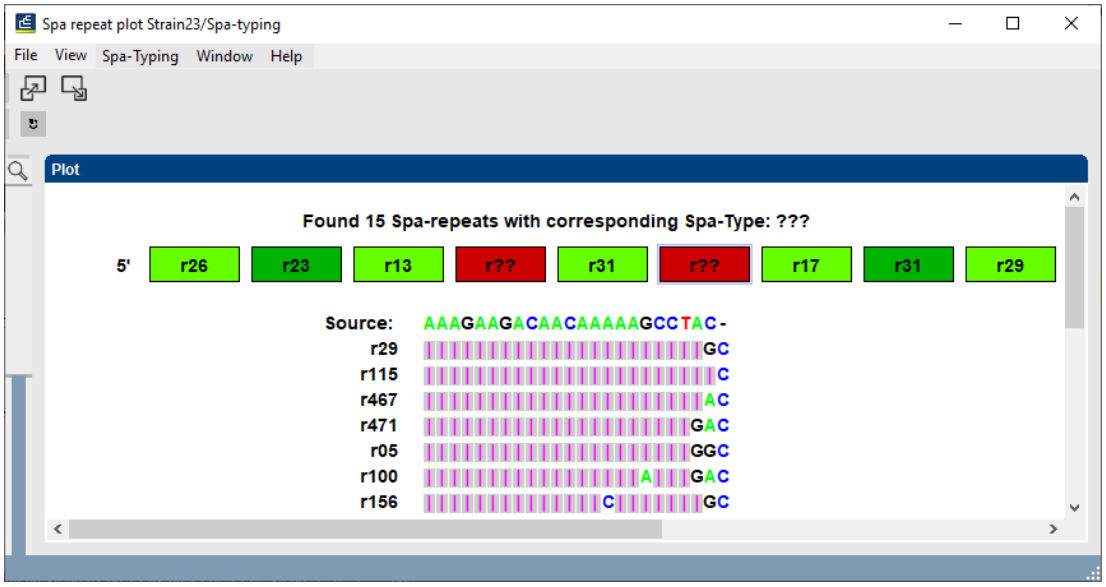


Figure 4.6: The repeat plot with editing suggestions for the second unknown repeat.

3.2 Click on the second unknown red "r??" repeat. A table is displayed with suggestions to edit the sequence. In the left column, the repeat is shown.

3.3 Use the zoom functions and (***View*** > ***Zoom in*** and ***View*** > ***Zoom out***) to obtain the best view of the plot.

Editing the sequence as suggested by the first row will give repeat "r29" (see Figure 4.6). Looking at this position in the *Contig assembly* window gives additional information about the missing base: in the chromatogram of the forward sequence, there is a missing "G".

3.4 Place the cursor on the gap in the trace sequence and type "G". The consensus sequence is automatically updated (see Figure 4.7 (b)).

3.5 Select the sequence you have edited (click on any position on the sequence, in the chromatogram or on the overview) and call ***File*** > ***Sequence information...*** (, **Ctrl+I**). This brings up the *Sequence information* dialog box, listing all base corrections that are made to the sequence. Press <***Exit***>.

3.6 Select ***Spa-Typing*** > ***Show repeats***.

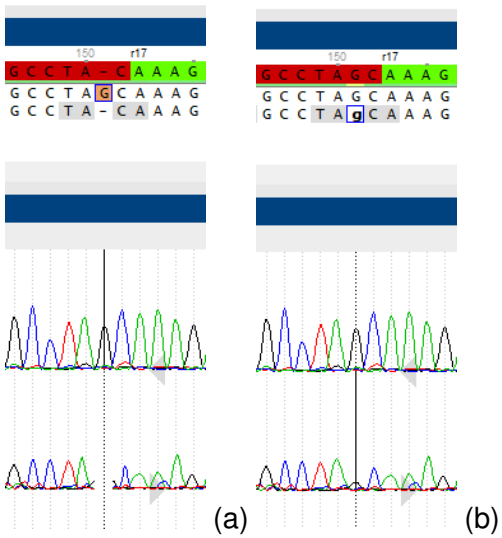


Figure 4.7: Missing peak in the chromatogram (a), Editing the trace sequence (b).

The repeat assignment in the consensus sequence is updated. The ”r29” repeat is displayed in green in the *Assembly view*.

3.7 Select ***Spa-Typing*** > ***Refresh*** in the repeat plot to update the information.

The corrected repeat is displayed in green.

3.8 Click on the remaining unknown repeat in the repeat plot.

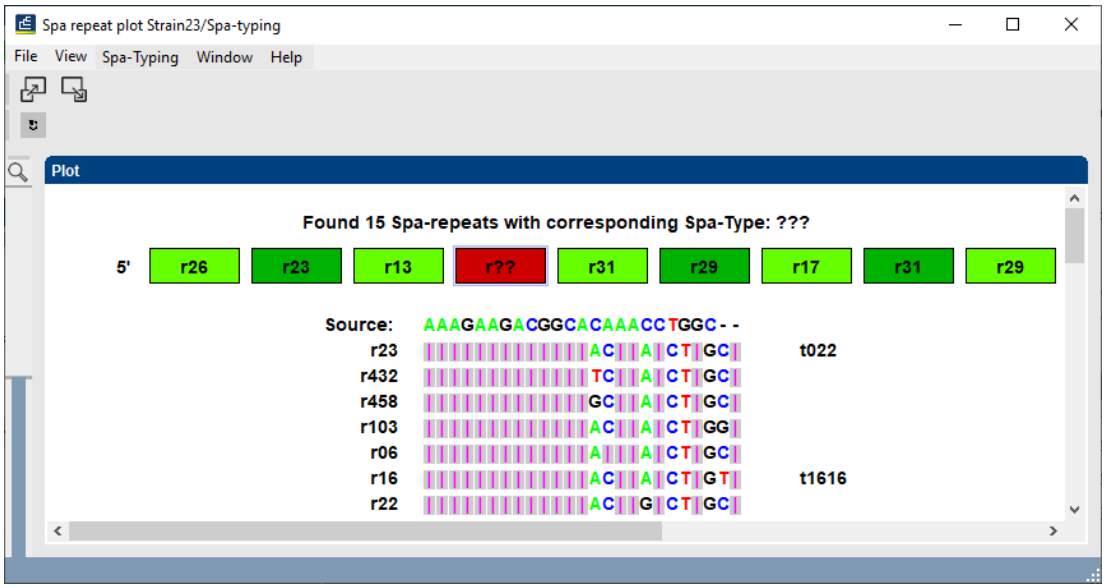


Figure 4.8: The repeat plot with editing suggestions for the remaining ”unknown” repeat.

The table with suggestions is displayed. In the left column, the repeat is shown. In the right column, the associated spa type is displayed (Figure 4.8). Editing the sequence as suggested by the first row will give repeat ”r23” and type ”t022”. Looking at this position in the *Contig assembly* window gives additional information: in the chromatogram of the forward sequence, there is a missing ”A” and based on the default ***Consensus determination*** parameters, this leads to a gap in the consensus sequence (see Figure 4.9 (a)).

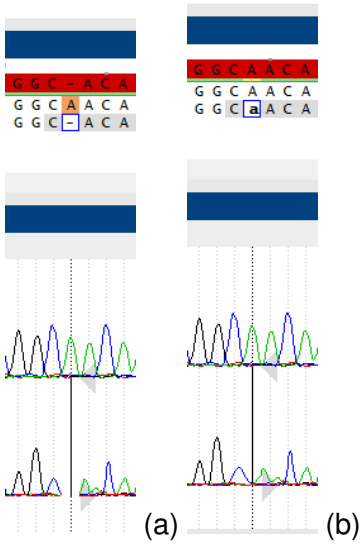


Figure 4.9: Missing peak in the chromatogram (a), Editing the trace sequence (b).

3.9 To insert the base in the trace sequence, place the cursor on the gap in the trace sequence and type ”A”.

The consensus sequence is automatically updated (see Figure 4.9 (b)).

3.10 Select ***Spa-Typing*** > ***Show repeats***.

The repeat assignment in the consensus sequence is updated. All repeats are now displayed in green in the *Assembly view*.

3.11 Select ***Spa-Typing*** > ***Refresh*** in the repeat plot to update the information.

3.12 To copy the repeat plot to the clipboard, select ***File*** > ***Copy to clipboard***.

3.13 The plot can be printed with ***File*** > ***Print***.

3.14 Close the *Repeat plot window* with ***File*** > ***Exit***.

The two warning messages (***Inconsistency in alignment at position 218*** and ***Inconsistency in alignment at position 160***) are checked and corrected for **Strain23**).

The plugin will not take into account unresolved bases in the consensus sequence when looking for spa repeats. Make sure no unresolved bases are present in the consensus sequence when looking for repeats.

The status of a contig project is set to **ERROR** if unresolved bases are detected in the consensus sequence.

4.4 Changing the status of error (and warning) messages

4.4.1 Principles

Only for those entries that have a green (= **OK** or **Solved**) or orange (= **Warning**) status, Spa types can be assigned.

- It is recommended to check the *warning* messages and solve them if needed. Since Spa types can be assigned to entries that have a Warning status, it is not required to change the status to ”Solved”.
- *Errors* need to be checked in the *Contig assembly* window and solved. Since Spa types cannot be assigned to entries that have an Error status, it is required to change the status to ”Solved” after having solved all errors in Assembler.

4.4.2 Option1: Changing the status in Assembler

4.1 Select ***Batch sequence assembly*** > ***Set report to solved, save and close*** (**Ctrl+Shift+S**) in the *Contig assembly* window.

The corresponding key/experiment cell in the overview *Batch sequence assembly report window* window is updated and displayed in green. The status ”Solved” is displayed in the key/experiment field.

4.4.3 Option2: Changing the status in the Detailed report window

4.2 Open the *Contig assembly* window for the entry with key **Strain5** by double-clicking on one of the two warning messages in the *Details* panel of the *Batch sequence assembly report window* window reporting an ***Inconsistency in alignment***.

The contig is shown in the *Contig assembly* window, with the corresponding position in focus.

4.3 Select ***Spa-Typing*** > ***Show repeats***.

The start and stop positions and 10 known repeats are detected.

4.4 Make sure the *Aligned traces* panel is selected and use the zoom sliders or the zoom buttons to obtain an optimal view of the curves.

If you look at the chromatograms at position 395 and 397, false peaks introduce a ”G” at position 395 and an ”A” at position 397 in the reverse sequence. Gaps in the forward sequence at these positions result in two gaps in the consensus sequence (see Figure 4.10). These gaps are the reason why two warning messages are reported for this contig project. We could delete these false base callings in the forward sequence, resulting in the removal of the gaps in the consensus sequence, but since we have allowed gaps to be present in the consensus sequence when searching for repeats and signatures in the source sequence, these gaps do not interfere with the analysis, and so we do not need to edit the sequence.

4.5 Select ***File*** > ***Save*** ( , **Ctrl+S**) and ***File*** > ***Exit*** to close the *Contig assembly* window.

4.6 In the *Batch sequence assembly report window* window, select ***Details*** > ***Set all messages to solved*** (**Ctrl+S**).

The corresponding key/experiment cell in the *Overview* panel is updated and displayed in green. The status ”solved” is displayed in the cell and in the ***Status*** column of the *Details* panel (see Figure 4.11).

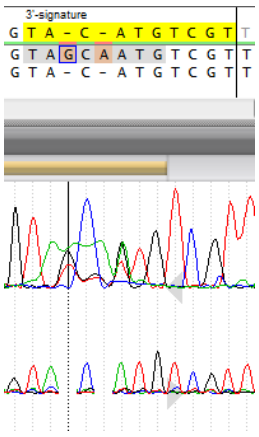


Figure 4.10: False peaks.

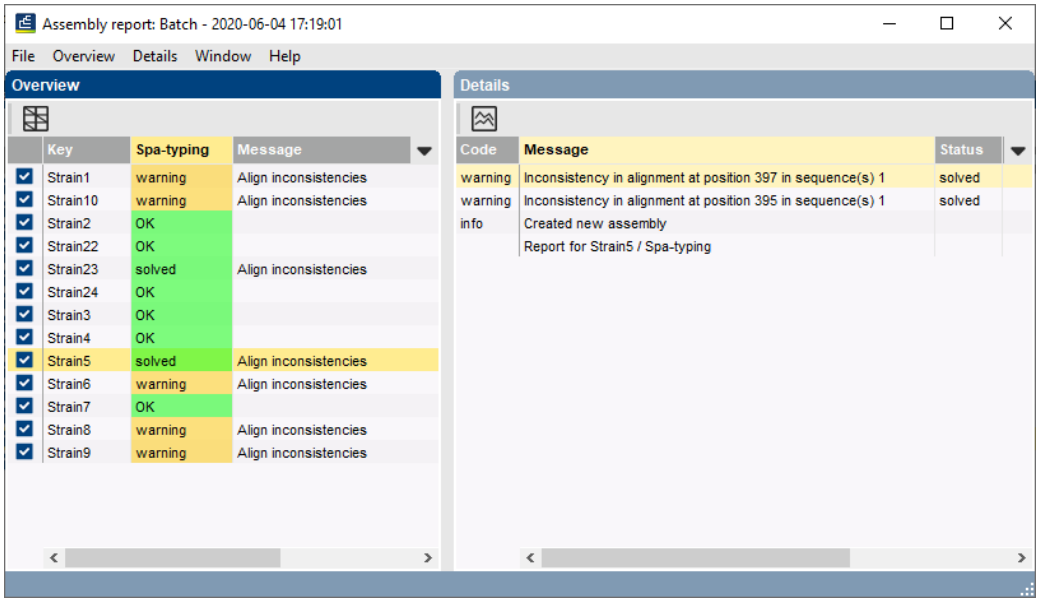


Figure 4.11: Solve errors/warnings.

Chapter 5

Spa-Typing in BIONUMERICS

5.1 Selections in the main window

In the *Main* window, a **Spa-typing** experiment is present for each contig project (see colored dot in the **Spa-typing** column in the *Experiment presence* panel).

Screening for spa repeats and types can be done for all entries present in the database, or for any selection of entries in database.

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**).

5.2 Assigning Spa types

5.2.1 Principles

2.1 Make a selection in the *Main* window.

2.2 Select ***Spa-Typing*** > ***Assign Spa types*** in the *Main* window.

The *Find Spa types* dialog box pops up (see Figure 5.1).

The ***Include Kreiswirth notation*** and ***Include clonal complexes*** check boxes are only shown if their corresponding information fields are present in the database (see Figure 1.5).

2.3 Press <***OK***>.

If no selection is present in the database, the software will display a message asking you if you wish to run the tool on the complete database.

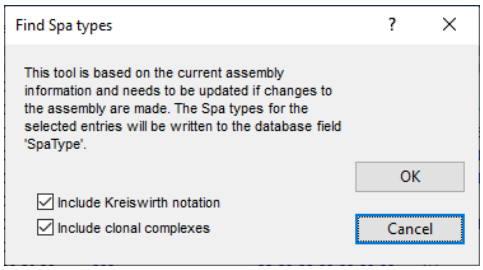


Figure 5.1: The *Find Spa types* dialog box.

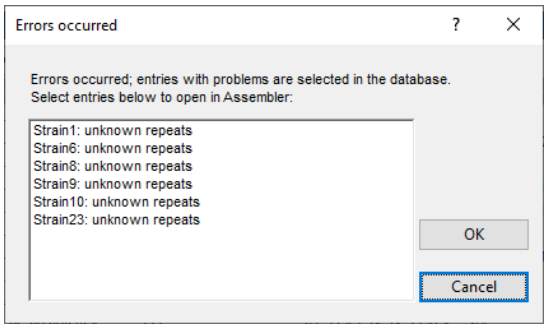


Figure 5.2: The *Errors occurred* dialog box.

If entries are detected with sequence assembly problems or unknown repeats, the *Errors occurred* dialog box pops up, listing all these entries with one of the following error messages:

- ***Unknown repeats***: One or more unknown repeats are detected in the consensus sequence.
- ***Problems with assembly***: The status box in the *Overview report window* reports an error message (= red status box). Spa types can only be assigned to entries that have a green (= **OK** or **Solved**) or orange (= **Warning**) status.

Entries can be selected and their assemblies can be opened in Assembler.

All entries with sequence assembly problems or unknown repeats are selected.

The *Spa typing plugin* uses a 2-step approach when the command ***Spa-Typing*** > ***Assign Spa types*** is selected:

5.2.2 Step 1: The assembly is screened for repeats

The repeat succession is displayed in the database information field that holds the repeat succession information (default name: **RepeatSuccession**, see Figure 1.5). The repeat succession is stored in the database information field that holds the repeat succession information *and* in the character type **Spa-repsuc**. If the ***Include Kreiswirth notation*** is checked in the *Find Spa types* dialog box, the Kreiswirth information is stored in the database information field that holds the Kreiswirth notation if this information field is present in the database.

2.4 Click on the colored dot in the **Spa-repsuc** column of the *Experiment presence* panel to open the character *Experiment card* window for an entry (see Figure 5.3).

2.5 Close the experiment card by clicking in the small triangle-shaped button in the left upper corner.

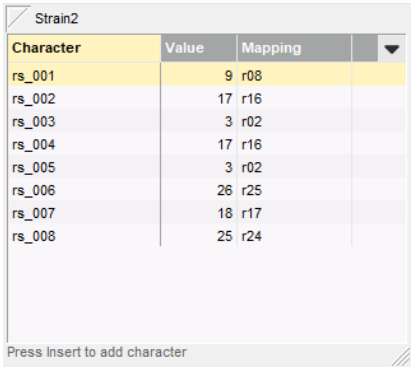


Figure 5.3: The Spa-repsuc character card, displaying the repeat succession in the ***Mapping*** column.

When a repeat sequence does not match one of the repeats in the database, or when a IUPAC code is present in the consensus sequence, a "??" is displayed in the **RepeatSuccession** information field and in the ***Mapping*** column of the character card. In the Kreiswirth information field - if present in the database - the text "NA" (Not Available) is displayed.

When a sequence is found that is too short or too long to be considered as a repeat sequence, an asterisk (*) is displayed in the **RepeatSuccession** information field and in the ***Mapping*** column of the character card. In the Kreiswirth information field - if present in the database - the text "NA" (Not Available) is displayed.

When no repeats are found, no information is written in the repeat succession and Kreiswirth information fields.

5.2.3 Step 2: Repeat type (if available) is assigned to each selected entry

The Spa type is displayed in the information field that holds the Spa Type information (default name: **SpaType**, see Figure 1.5).

The Spa type is denoted as "???" if the repeat succession is incomplete. When the repeat information is currently not linked to a Spa type in the database, "Unknown" is displayed in the spa type information field. If no repeats are found, "NA"(Not Available) is displayed.

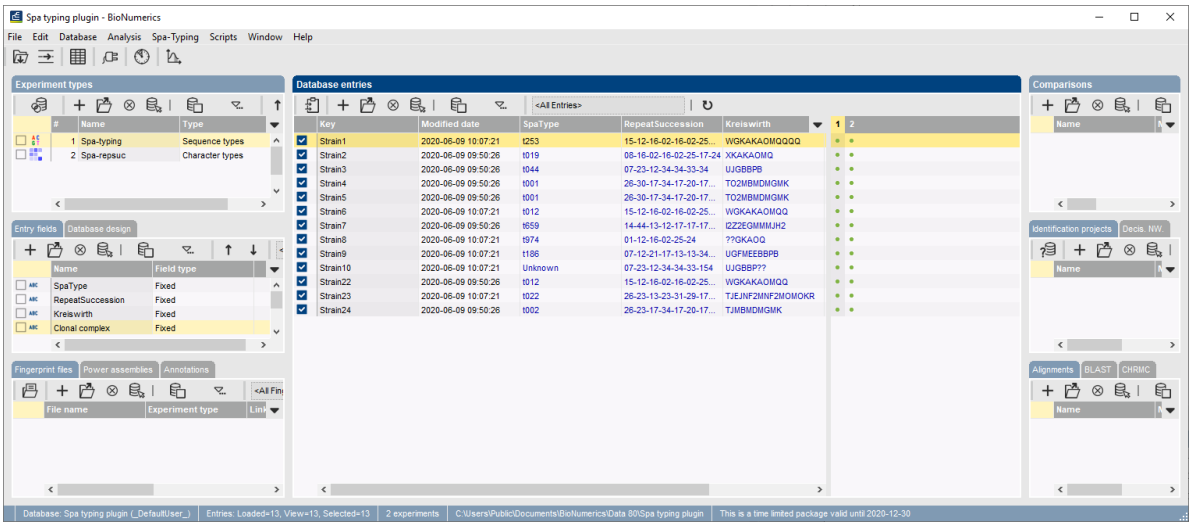


Figure 5.4: The *Main* window after repeat and type assignment.

Chapter 6

Cluster analysis of Spa types

6.1 Introduction

In this chapter, we are going to take a look at the evolutionary relationship between the Spa sequences by means of the construction of a dendrogram and a minimum spanning tree.

The *Spa typing plugin* uses a multi-step approach for this cluster analysis.

- The plugin uses an algorithm based on a DSI model [1] for the pairwise alignment of the Spa repeats. This model considers three mutational events: Duplication of tandem repeats, Substitutions and Indels.
- Next, the cost matrix is used to correct for the evolutionary distances between the repeats.

Taking these costs into account, the output of the DSI model is a similarity matrix. From this similarity matrix a dendrogram and/or a minimum spanning tree can be constructed.

6.2 The Comparison window

2.1 For this exercise, make sure all entries are selected in the *Main* window (**Ctrl+A**).

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.

2.3 Drag the separator lines between the panels to the left or to the right, in order to divide the space among the panels optimally.

2.4 Move the panels by clicking in the header of a panel and - while keeping the mouse button pressed - dragging it to another location in the *Comparison* window.

The character type **Spa-repsuc** is created upon installation of the *Spa typing plugin* and displayed in the *Experiments* panel. The repeat information stored in the associated character type will be used when using the clustering tools. The repeat succession stored in the associated repeat information field is only used when no repeat information is present in the associated character type.

2.5 Click on the eye button ( ) of the character type **Spa-repsuc** in the *Experiments* panel.

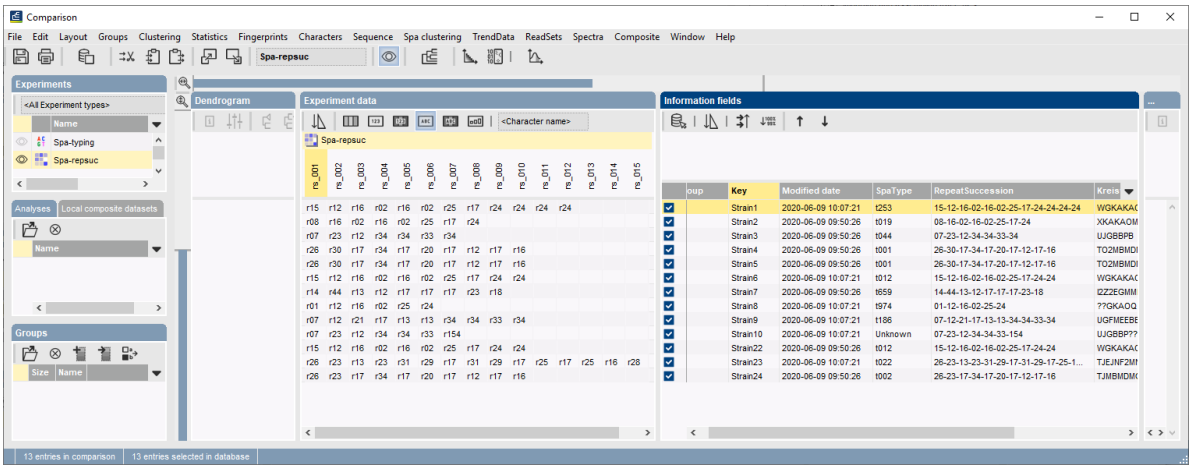


Figure 6.1: The *Comparison* window.

The pattern images are displayed in the *Experiment data* panel. Initially, the character values are displayed as colors according to the color scale defined for each character (see the reference manual for more information).

2.6 Select ***Characters*** > ***Show mappings*** ( ) or ***Characters*** > ***Show mappings+colors*** ( ) to display the mapped name for each character value (see Figure 6.1).

6.3 Creating a cost matrix

In the *Spa typing plugin*, there is a default binary cost matrix available for the calculation of the dendrogram, consisting of two states: a match between the repeats and no match.

3.1 Select ***Spa clustering*** > ***Cost matrices*** in the *Comparison* window for the creation of your own cost matrix.

The *Cost matrices* dialog box appears (see Figure 6.2).

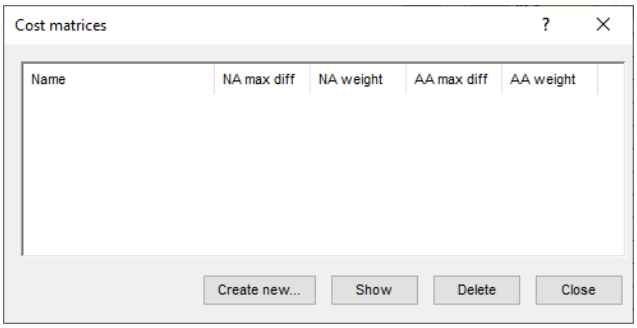


Figure 6.2: The *Cost matrices* dialog box.

The *Cost matrices* dialog box displays all cost matrices defined by the user (initially empty).

Selecting <***Create new***> displays the *Create Spa cost matrix* dialog box (see Figure 6.3).

You can define a ***Name*** for the cost matrix and set the costs for nucleotides and amino acids.

- ***Maximum number of differences***: defines the maximum number of differences in nucleotides/amino acids between two repeats. The default is 5 and there is a gradual cost

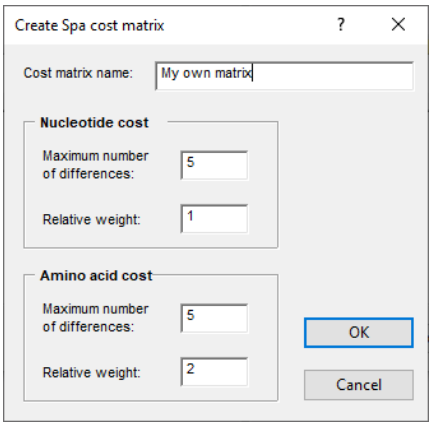


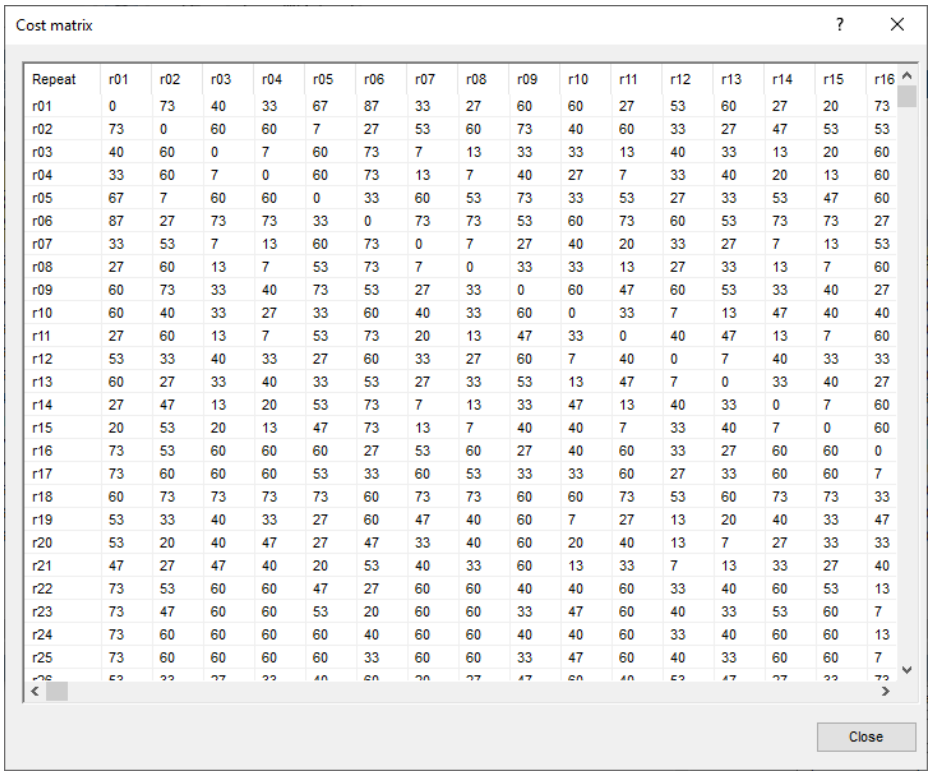
Figure 6.3: The *Create Spa cost matrix* dialog box.

between 0 and 5 mismatches. Differences larger than 5 will get 100% of the cost as well.

- ***Relative weight***: defines the relative weight between the nucleotides and the amino acids. The settings in Figure 6.3 penalize a change in an amino acid twice as much as a change in a nucleotide.

3.2 Select <***Create new***>, specify a ***Cost matrix name***, leave the settings unaltered and press <***OK***>.

This calls the *Cost matrix* dialog box (see Figure 6.4).



Repeat	r01	r02	r03	r04	r05	r06	r07	r08	r09	r10	r11	r12	r13	r14	r15	r16
r01	0	73	40	33	67	87	33	27	60	60	27	53	60	27	20	73
r02	73	0	60	60	7	27	53	60	73	40	60	33	27	47	53	53
r03	40	60	0	7	60	73	7	13	33	33	13	40	33	13	20	60
r04	33	60	7	0	60	73	13	7	40	27	7	33	40	20	13	60
r05	67	7	60	60	0	33	60	53	73	33	53	27	33	53	47	60
r06	87	27	73	73	33	0	73	73	53	60	73	60	53	73	73	27
r07	33	53	7	13	60	73	0	7	27	40	20	33	27	7	13	53
r08	27	60	13	7	53	73	7	0	33	33	13	27	33	13	7	60
r09	60	73	33	40	73	53	27	33	0	60	47	60	53	33	40	27
r10	60	40	33	27	33	60	40	33	60	0	33	7	13	47	40	40
r11	27	60	13	7	53	73	20	13	47	33	0	40	47	13	7	60
r12	53	33	40	33	27	60	33	27	60	7	40	0	7	40	33	33
r13	60	27	33	40	33	53	27	33	53	13	47	7	0	33	40	27
r14	27	47	13	20	53	73	7	13	33	47	13	40	33	0	7	60
r15	20	53	20	13	47	73	13	7	40	40	7	33	40	7	0	60
r16	73	53	60	60	60	27	53	60	27	40	60	33	27	60	60	0
r17	73	60	60	60	53	33	60	53	33	33	60	27	33	60	60	7
r18	60	73	73	73	73	60	73	73	60	60	73	53	60	73	73	33
r19	53	33	40	33	27	60	47	40	60	7	27	13	20	40	33	47
r20	53	20	40	47	27	47	33	40	60	20	40	13	7	27	33	33
r21	47	27	47	40	20	53	40	33	60	13	33	7	13	33	27	40
r22	73	53	60	60	47	27	60	60	40	40	60	33	40	60	53	13
r23	73	47	60	60	53	20	60	60	33	47	60	40	33	53	60	7
r24	73	60	60	60	60	40	60	60	40	40	60	33	40	60	60	13
r25	73	60	60	60	60	33	60	60	33	47	60	40	33	60	60	7

Figure 6.4: The *Cost matrix* dialog box.

The cost matrix is calculated and shown. The higher the costs, the more distantly related the repeats are. Press <***Close***> to close the *Cost matrix* dialog box.

A selected cost matrix is removed from the list in the *Cost matrices* dialog box with <***Delete***>.

The cost matrix is shown in the *Cost matrix* dialog box when pressing the <***Show***> button.

The *Cost matrices* dialog box can be closed with <***Close***>.

6.4 Cluster analysis settings

4.1 Select ***Spa clustering*** > ***Cluster Spa types*** in the *Comparison* window.

The *Spa Clustering* dialog box appears (see Figure 6.5).

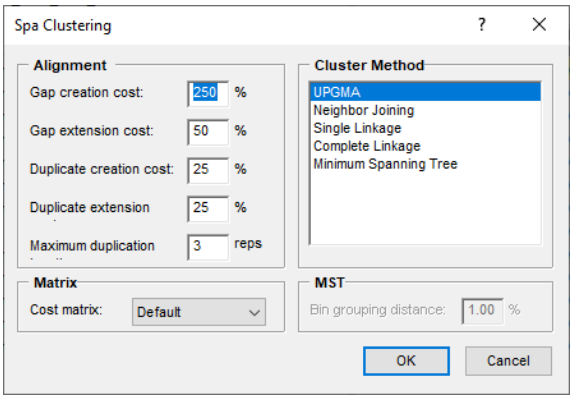


Figure 6.5: The *Spa Clustering* dialog box.

Following settings can be specified in the *Spa Clustering* dialog box:

Alignment settings:

- ***Gap creation cost***: specifies the cost for the introduction of a single gap in one of the repeats (in %).
- ***Gap extension cost***: defines the cost for the extension of a created gap (in %).
- ***Duplicate creation cost***: gives the cost for the duplication of a repeat (in %).
- ***Duplicate extension***: defines the cost for the extension of a duplicated repeat (in %).
- ***Maximum duplication length***: defines the maximum number of neighboring repeats that are taking into account to create a duplicate from.

Matrix:

In the *Matrix panel*, the default cost matrix or a custom cost matrix can be selected from the drop-down menu (see 6.3 for the creation of a cost matrix).

Cluster Method:

In the upper right box, five cluster methods are listed: ***Minimum spanning tree***, ***UPGMA***, ***Neighbor Joining***, ***Single Linkage***, and ***Complete Linkage***.

An additional setting called ***Distance bin size*** is displayed in the ***MST panel*** when the ***Minimum spanning tree*** option is checked. Based on this setting, the software creates bins of certain distance intervals, that are converted into distance units. When for example the distance bin size is set to 1%, two entries having a similarity of 99.6% will have a distance of 0 (interval 100%-99% =

distance 0). Two entries that have a similarity of 98.7% will have a distance of 1 (interval 99%-98% = distance 1). The default setting is 1%.

In this example, we will create a minimum spanning tree (see 6.5) and a UPGMA dendrogram (see 6.6).

6.5 Minimum spanning tree

Minimum spanning trees are trees calculated from a distance matrix and possess the property of having a total branch length that is as small as possible. A MST chooses the sample with the highest number of related samples as the root node, and derives the other samples from this node. This may result in trees with star-like branches and allows for a correct classification of population systems that have a strong mutational or recombinational rate.

5.1 Select ***Spa Clustering*** > ***Cluster Spa types*** in the *Comparison* window and select ***Minimum Spanning Tree*** in the *Cluster Method panel* (see Figure 6.5).

An additional setting called ***Distance bin size*** is displayed in the ***MST panel***. Based on this setting, the software creates bins of certain distance intervals, that are converted into distance units. When for example the distance bin size is set to 1%, two entries having a similarity of 99.6% will have a distance of 0 (interval 100%-99% = distance 0). Two entries that have a similarity of 98.7% will have a distance of 1 (interval 99%-98% = distance 1). The default setting is 1%.

5.2 Leave the settings unaltered and press <***OK***>.

The *Cluster analysis* window pops up (see Figure 6.6). The *Network panel* displays the minimum spanning tree, the upper right panel (*Entry list*) displays the entries that are present in the tree. The *Selection entry list* lists the entries that are present in the selected node(s).

5.3 Select a node or branch by clicking on them, or several nodes/branches by holding the **Shift**-key while clicking.

As an exercise we will change some display settings. More detailed information about the *Cluster analysis* window can be found in the reference manual.

5.4 Choose ***Display*** > ***Display settings*** to open the *Display settings* dialog box.

5.5 In the *Node labels and sizes tab*, select ***Show node labels*** and select ***SpaType*** from the list.

5.6 In the *Node colors tab*, select ***Number of entries*** from the drop-down list.

5.7 In the *Branch styles tab*, select ***branch length*** from the drop-down list.

5.8 In the *Branch labels and sizes tab*, select ***Show branch labels*** and ***branch length***.

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

The *Cluster analysis* window should now look like Figure 6.6.

5.10 In the *Cluster analysis* window, select ***Display*** > ***Zoom to fit*** to optimize the view of the tree in the current window.

5.11 Close the *Cluster analysis* window.

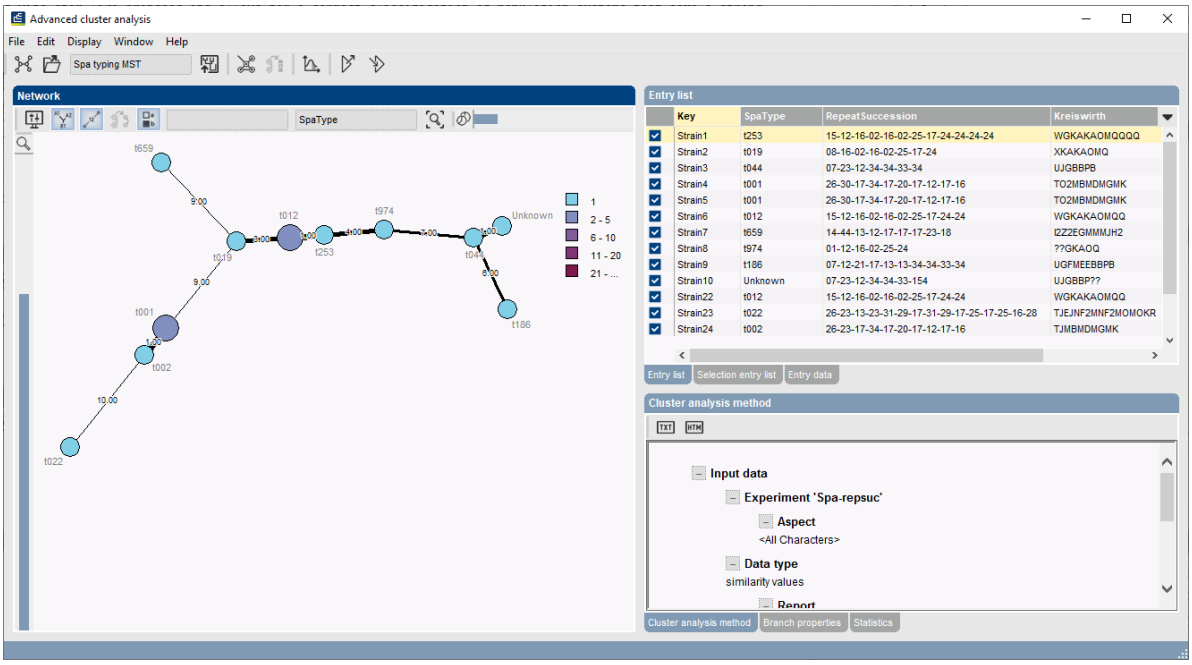


Figure 6.6: The *Cluster analysis* window.

6.6 Cluster analysis sensu stricto

Cluster analysis *sensu stricto* is based upon the similarity matrix and a subsequent algorithm for calculating bifurcating dendrograms to cluster the entries. In the *Spa typing plugin* you can choose between the following four methods: Unweighted Pair Group Method using Arithmetic averages (***UPGMA***), the ***Neighbor Joining*** method and two variants of UPGMA: ***Single linkage*** and ***Complete linkage*** (see Figure 6.5).

6.1 In the *Comparison* window, choose ***Spa clustering*** > ***Cluster Spa types***.

6.2 Select ***UPGMA***, use the default alignment settings and default cost matrix and press <***OK***>.

The dendrogram is shown in the *Comparison* window (see Figure 6.7).

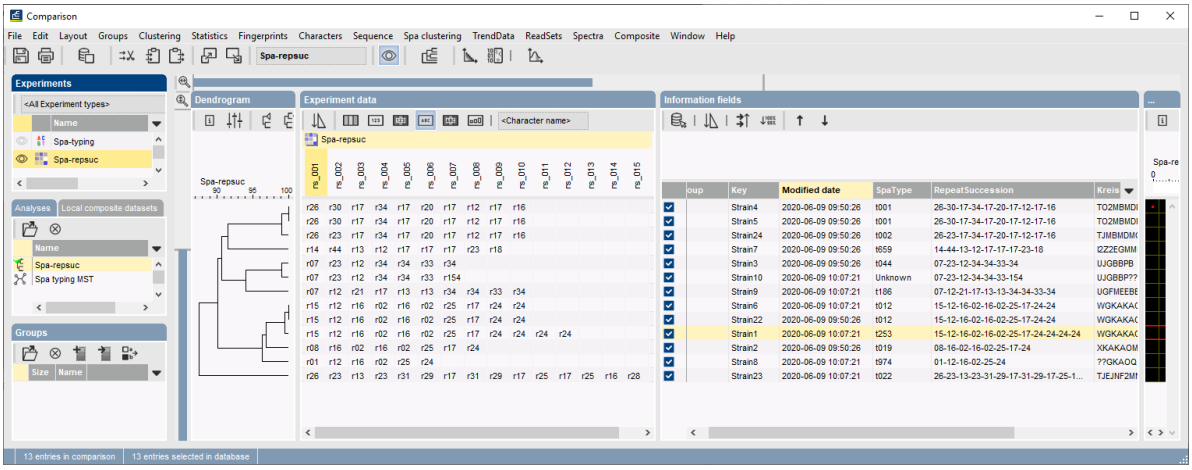


Figure 6.7: The *Comparison* window with a dendrogram and a similarity matrix.

6.3 Click on the dendrogram to place a cursor on any node or tip (where a branch ends in an individual entry). The average similarity at the cursor's place is shown in the upper part of the *Experiment data* panel. You can move the cursor with the arrow keys.

More detailed information about the dendrogram display settings can be found in the reference manual.

6.4 Save and close the *Comparison* window.

Chapter 7

Matching Spa types

7.1 Selections in the main window

One or more selected Spa types can be matched (identified) against the complete database, all Spa types, or a selection in the database.

1.1 As an exercise, select a few entries in the *Main* window (e.g. **Strain22**, **Strain23**, and **Strain 24**).

7.2 Matching Spa types

2.1 Call the *Spa matching* dialog box with ***Spa-Typing*** > ***Match Spa types***.

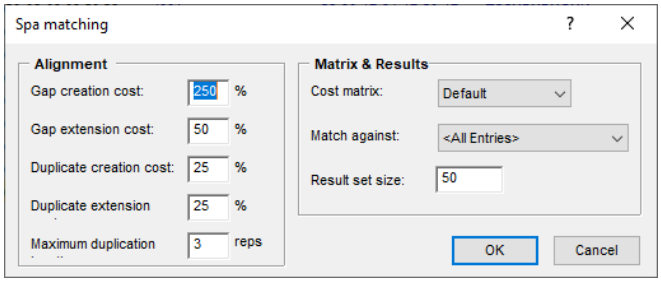


Figure 7.1: The *Spa matching* dialog box with the default settings.

In the *Spa matching* dialog box, following settings can be specified:

Alignment settings:

- ***Gap creation cost***: specifies the cost for the introduction of a single gap in one of the repeats (in %).
- ***Gap extension cost***: defines the cost for the extension of a created gap (in %).
- ***Duplicate creation cost***: gives the cost for the duplication of a repeat (in %).
- ***Duplicate extension***: defines the cost for the extension of a duplicated repeat (in %).
- ***Maximum duplication length***: defines the maximum number of neighboring repeats that are taking into account to create a duplicate from.

Matrix & Results:

- ***Cost matrix***: The drop-down menu lists the default cost matrix and the user-defined cost matrices (if created).
- ***Match against***: The selection can be matched against all entries in the database (<***All Entries***>), all entries of which the currently logged-in user is the owner (<***My Entries***>), all entries currently loaded into memory (<***Loaded Entries***>), all selected entries (<***Selected Entries***>), or all known types (<***All Spa types***>).
- ***Result set size***: Defines the number of best matches that are shown in the detailed report.

2.2 For this exercise, choose <***All Entries***> from the ***Match against*** menu, leave all other settings at their defaults and press <***OK***>.

The program tries to find the best matches for the selected entries based on their repeats. The *Spa matching window* appears (see Figure 7.2).

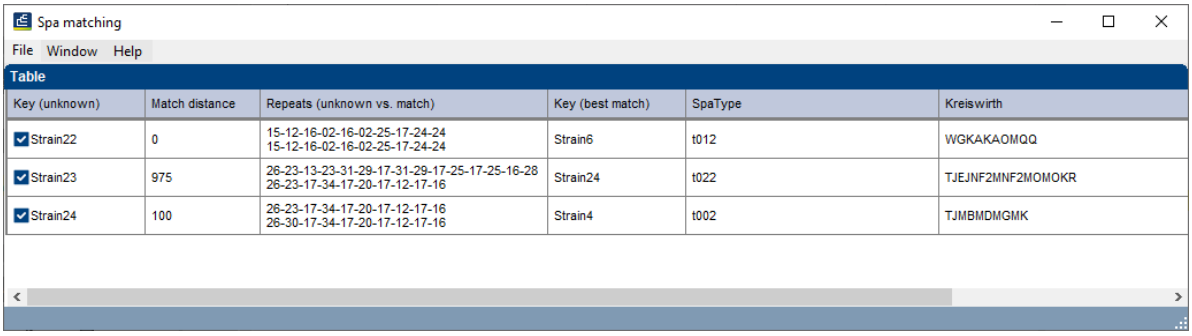


Figure 7.2: The *Spa matching window*.

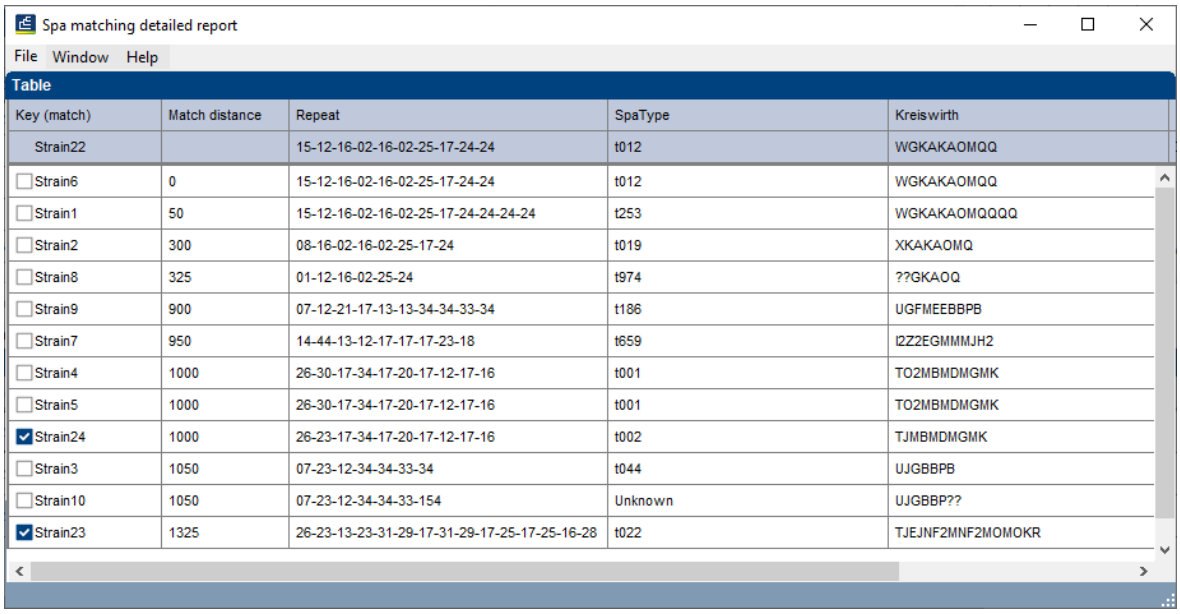
the repeat information stored in the associated character type will be used when matching entries. The repeat succession stored in the associated repeat information field is only used when no repeat information is present in the associated character type.

- In the first column, the keys of the selected ”unknown” entries are shown.
- The distance between the selected entry and its match is displayed in the second column. The smaller the value, the better the match with ”0” being an exact match.
- The repeats of the selected entries and their matches are shown in the third column.
- The fourth column displays the best matching entry.
- In the last column(s), the entry field content of the unknown entry is listed.

2.3 Double-click on an entry in the *Spa matching window* (e.g. entry with key **Strain22**).

A detailed report pops up (see Figure 7.3). The best matching entries are shown in descending order.

In both report windows, you can select or unselect entries by pressing the **Ctrl-** or **Shift-** key while holding the left mouse button.



Key (match)	Match distance	Repeat	SpaType	Kreiswirth
Strain22		15-12-16-02-16-02-25-17-24-24	t012	WGKAKAOMQQ
Strain6	0	15-12-16-02-16-02-25-17-24-24	t012	WGKAKAOMQQ
Strain1	50	15-12-16-02-16-02-25-17-24-24-24-24	t253	WGKAKAOMQQQQ
Strain2	300	08-16-02-16-02-25-17-24	t019	XKAKAOMQ
Strain8	325	01-12-16-02-25-24	t974	??GKAOQ
Strain9	900	07-12-21-17-13-13-34-34-33-34	t186	UGFMEEBBPB
Strain7	950	14-44-13-12-17-17-17-23-18	t659	I2Z2EGMMMJH2
Strain4	1000	26-30-17-34-17-20-17-12-17-16	t001	TO2MBMDMGMK
Strain5	1000	26-30-17-34-17-20-17-12-17-16	t001	TO2MBMDMGMK
Strain24	1000	26-23-17-34-17-20-17-12-17-16	t002	TJMBMDMGMK
Strain3	1050	07-23-12-34-34-33-34	t044	UJGBBPB
Strain10	1050	07-23-12-34-34-33-154	Unknown	UJGBBP??
Strain23	1325	26-23-13-23-31-29-17-31-29-17-25-17-25-16-28	t022	TJEJNF2MNF2MOMOKR

Figure 7.3: Detailed report of the *Spa matching window*.

Chapter 8

Synchronizing with SpaServer

8.1 SpaServer information fields

In BIONUMERICS it is possible to submit new Spa types to the online SpaServer via a synchronization process. Spa data can only be submitted to the online SpaServer, if all mandatory SpaServer strain information is provided when uploading the information to the SpaServer.

Mandatory SpaServer strain information includes: **isolation date** (YYYY-MM-DD), **country**, **MRSA/MSSA** (MRSA, MSSA), and **origin** (person, animal, environment, unknown). More information can be found on the *Submission* page of the SpaServer website.

A number of information fields are automatically created when a new database is created and after installation of the *Spa typing plugin* (see Figure 1.7).

In addition, extra information fields can be added to the *Database entries* panel with ***Edit*** > ***Information fields*** > ***Add information field...***. This command can also be accessed by right-clicking in the information toolbar of the *Database entries* panel.

1.1 Add information fields to the database for the storage of all mandatory SpaServer strain information: isolation date, country, MRSA/MSSA, and origin (see Figure 8.3 for an example).

1.2 Optionally, add information fields to the database for the storage of additional (not mandatory) strain information (e.g. City, ZIP,. . .).

Strain information can be entered in the database in several ways:

- Importing information stored outside BIONUMERICS (e.g. in a text file or an ODBC-compatible source) with the import routines available via ***File*** > ***Import...*** (, **Ctrl+I**).
- Entering information using the *Entry* window (double-click on an information field to call this window).
- Editing information directly by clicking twice on an information field.

Detailed information on each of these options can be found in the reference manual. In this section only the second option will be illustrated.

1.3 Double-click on a database entry to open the *Entry* window. Right-clicking on the entry, and selecting ***Open highlighted entry*** also opens this window.

In default configuration, the upper left panel of the *Entry* window shows the information fields. The upper right panel shows the available experiments for the entry (see Figure 8.1).

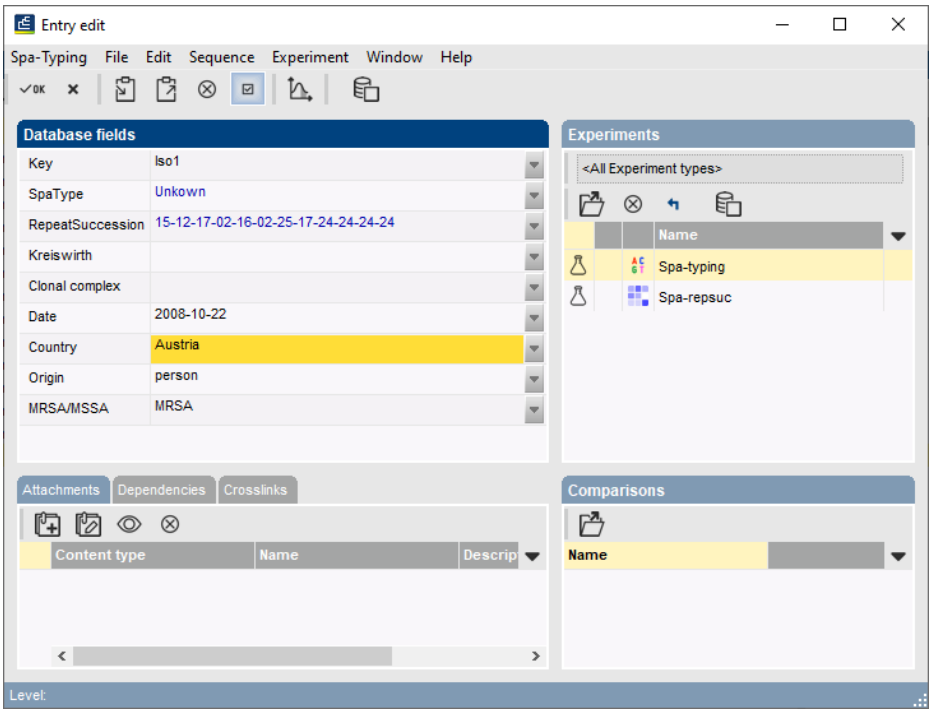


Figure 8.1: The *Entry* window.

1.4 Enter the information in the fields.

In the *SpaServer synchronization settings* dialog box (see 8.2), the BIONUMERICS information fields containing mandatory (and optional) strain information can be linked to the SpaServer information fields (see 8.2.3). When there is a link present between the BIONUMERICS information fields and the **MRSA/MSSA** and the **Origin** SpaServer information fields (see 8.2.3), the history lists for these BIONUMERICS information fields contain all possible online options (**Origin**: unknown, person, animal, environment; **MRSA/MSSA**: MRSA, MSSA). These history lists can be used to save time and work and to avoid typographical errors.

1.5 The history lists can be accessed by clicking the button on the right hand from the information field in the *Entry* window. A floating menu appears from which the correct information string can be selected (see Figure 8.2).

Figure 8.2: History list in the *Entry* window.

1.6 Press the **Enter**-key or select <***OK***> to close the *Entry* window. The information is stored in the database.

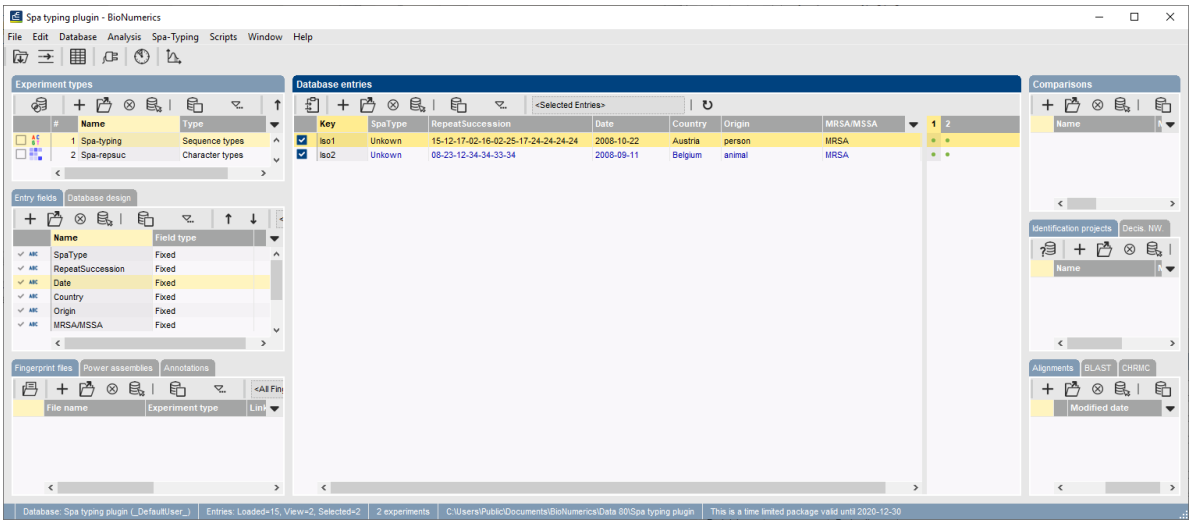


Figure 8.3: The *Main* window with information fields containing date, country, MRSA/MSSA and origin information.

8.2 SpaServer synchronization settings

8.2.1 Introduction

Before new Spa types can be submitted to the SpaServer, some synchronization settings need to be specified in BIONUMERICS. These settings can be accessed via the menu command ***Spa-Typing*** > ***SpaServer synchronization settings***.

2.1 Select ***Spa-Typing*** > ***SpaServer synchronization settings*** in the *Main* window.

This pops up the *SpaServer synchronization settings* dialog box (see Figure 8.4).

8.2.2 Add SpaServer users

Synchronization with the SpaServer is only possible if at least one registered SpaServer user is defined in the BIONUMERICS database. To edit or view the user settings of the user selected in the list *SpaServer users panel* press the <***Edit***> button. All user information of the selected user is deleted with the <***Delete button***>. To add the contact details of a registered SpaServer user to the database, select the <***Add***> button in the *SpaServer users panel*.

2.2 Press the <***Add***> button to call the *Add user* dialog box (see Figure 8.5).

The *Add user* dialog box prompts for the user ID of the new user.

2.3 Enter a user ID in the *Add user* dialog box and press <***OK***> to call the *Edit SpaServer user* dialog box (see Figure 8.6).

In the *Edit SpaServer user* dialog box, all information fields marked with an asterisk are mandatory fields.

To obtain a ***SeqNet.org release code***, please contact SeqNet.org.

A BIONUMERICS script is available that generates XML files of the certification trial data processed in BIONUMERICS. Please contact Applied Maths to obtain this script.

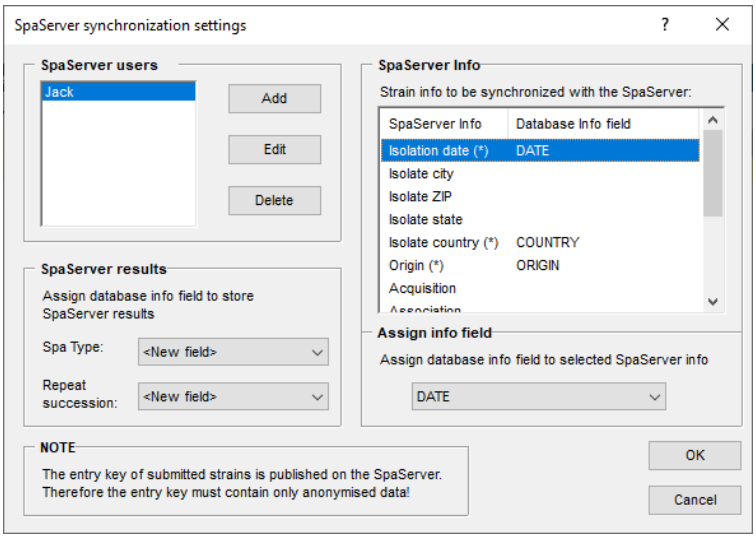


Figure 8.4: The *SpaServer synchronization settings* dialog box.

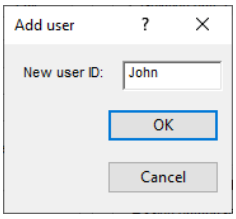


Figure 8.5: The *Add user* dialog box.

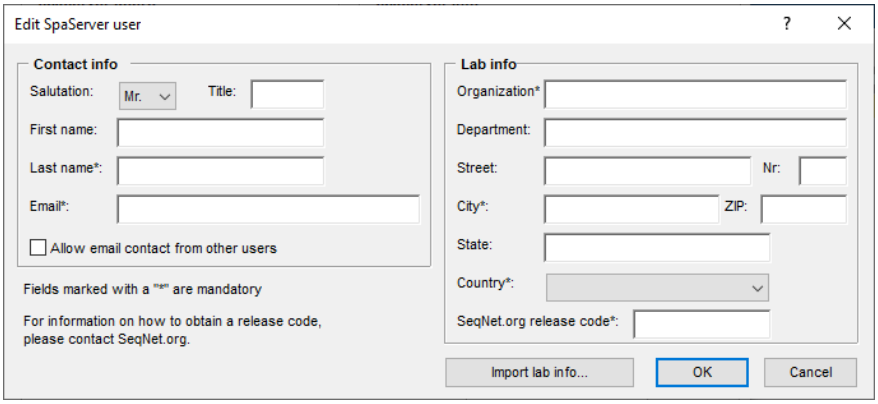


Figure 8.6: The *Edit SpaServer user* dialog box.

When there is at least one user defined in the list of SpaServer users, the option <***Import lab info***> becomes available when a new SpaServer user is added to the database. Pressing this button calls the *Import lab info* dialog box.

The Lab information of the selected user is copied to the *Lab info panel* of the new user when pressing <***OK***>.

2.4 Enter the user information in the *Edit SpaServer user* dialog box and press <***OK***> to add the SpaServer user to the database.

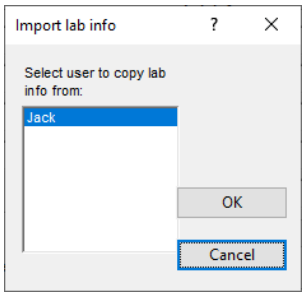


Figure 8.7: The *Import lab info* dialog box.

8.2.3 Link BIONUMERICS information fields to SpaServer fields

In the *SpaServer Info panel*, the SpaServer fields are listed in the *SpaServer info column*. SpaServer fields marked with an asterisk are mandatory fields, and need to be linked to one of the BIONUMERICS information fields. Fields without an asterisk are optional fields. The database information field selected in the *Assign info field panel* is displayed in the right column of the *SpaServer info panel*. All mandatory SpaServer fields need to be linked to their corresponding database information fields. If one or more mandatory SpaServer fields are not linked to a BIONUMERICS information field, the synchronization will fail.

The entry key is automatically linked to the SpaServer ID.

2.5 In the *SpaServer info panel*, select a SpaServer info field in the left column and select the corresponding database information field from the drop-down menu in the *Assign info field panel*.

2.6 Repeat the previous action for at least all mandatory SpaServer fields.

8.2.4 Store SpaServer results in BIONUMERICS database fields

BIONUMERICS information fields to store the SpaServer ***SpaType*** and ***Repeat succession*** results can be selected from two drop-down menus in the *SpaServer results panel*. An existing or new field can be selected from the list.

2.7 After having specified all settings in the *SpaServer synchronization settings* dialog box press <***OK***>.

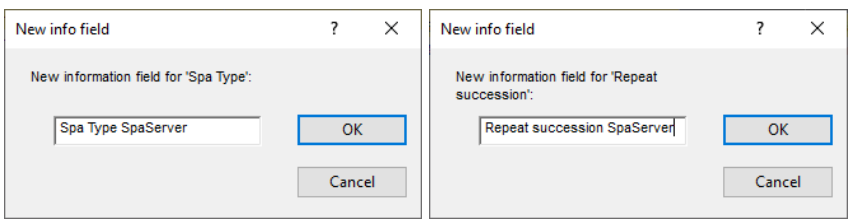


Figure 8.8: The *New info field* dialog box.

If the online results are assigned to new database information fields the *New info field* dialog box is displayed prompting for the information field name(s).

2.8 Specify a new information field name for the storage of the spa type and/or the repeat succession, or keep the default suggested name(s). Press <***OK***> to add the fields to the database.

8.3 Synchronizing with SpaServer (batch mode)

Spa data present in the BIONUMERICS database can be synchronized with the SpaServer in *batch*.

3.1 Select the entries you wish to synchronize with the SpaServer. To select entries use the **Ctrl-** and **Shift**-keys. Check boxes for selected entries are indicated as ☑.

3.2 Select ***Spa-Typing*** > ***Synchronize with SpaServer*** in the *Main* window.

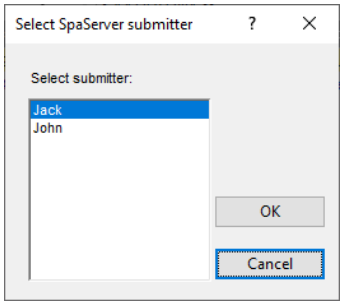


Figure 8.9: The *Select SpaServer submitter* dialog box.

If more than one SpaServer user is defined in the database, the *Select SpaServer submitter* dialog box pops up, listing all users defined in the *SpaServer synchronization settings* dialog box. Select the correct user from the list and press <***OK***>.

When there are no users defined in the database, an error message pops up (see Figure 8.10 (a)). Select ***Spa-Typing*** > ***SpaServer synchronization settings*** in the *Main* window, and press the <***Add***> button to add a spa user to the database (see 8.2).

When one or more mandatory SpaServer fields are not linked to a BIONUMERICS information field, an error message pops up (see Figure 8.10 (b)). When pressing <***OK***>, the *SpaServer synchronization settings* dialog box automatically pops up.

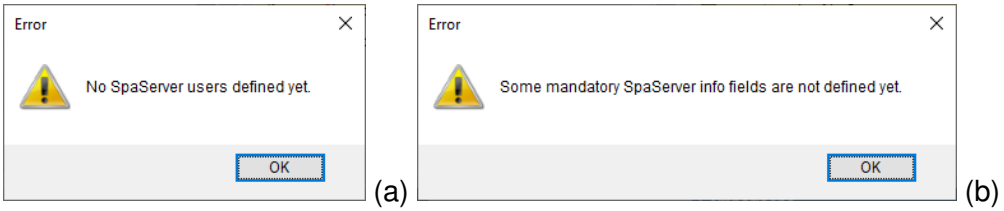


Figure 8.10: An error message pops up when (a) no SpaServer users are defined in the database,(b) when not all mandatory SpaServer fields are linked to a BIONUMERICS information field.

3.3 When all fields are correctly linked BIONUMERICS tries to submit the data to the Ridom/Seqnet SpaServer. The *SpaServer synchronization* dialog box is displayed (see Figure 8.11).

The *SpaServer synchronization* dialog box lists all selected entries. In the left column, the keys of the entries are shown. Depending on the information that BIONUMERICS has tried to submit to the online SpaServer, different messages are displayed in the **SpaServer response** and **BN error** columns in the *Report window*.

BN errors:

- ***The experiment "Spa-Typing" is not present***: no Spa-Typing experiment is defined for the

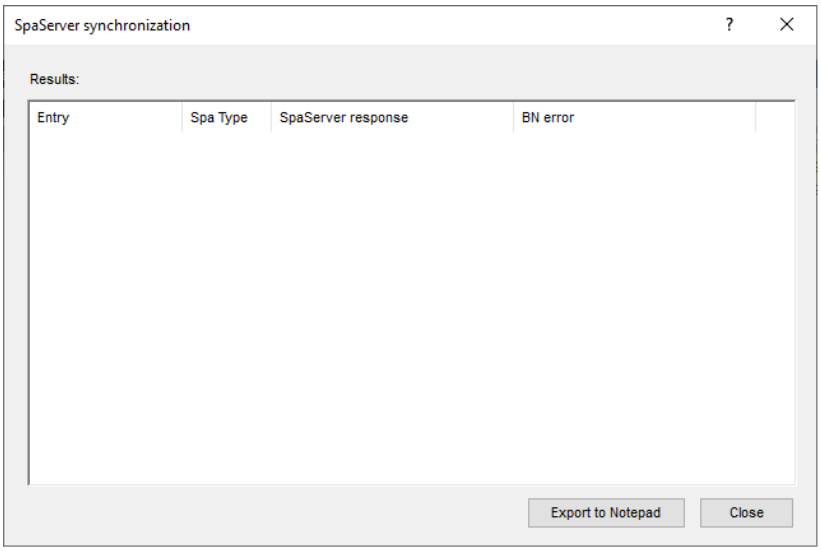


Figure 8.11: The *SpaServer synchronization* dialog box.

selected entry.

- ***Sequence is empty***: a Spa-Typing experiment is present for the selected entry, but this experiment does not contain an assembly.
- ***Assembly quality problem***: the assembly quality settings differ from the default settings that are required for the submission of new types to the SpaServer, and/or the number of active traces in the contig is not equal to 2, and/or one or both trimming patterns are not detected on the consensus sequence.
- ***Problems with assembly***: the status of the assembly is not set to solved.
- ***Invalid input in field "Name information field"***: the information field in the database does not contain the correct information.

SpaServer responses:

- ***New Spa type; SpaServer ID***: a new Spa type is submitted to the SpaServer. The new Spa type is shown in the **Spa Type** column (see Figure 8.11).
- ***Existing Spa type***: this message is displayed when an existing spa type is submitted to the SpaServer. The Spa type is shown in the second column.

Only the strain info of NEW Spa types are stored. Updating information fields of already existing types is not possible.

- ***The LabCode ****** is not valid!***: the SeqNet.org release code of SpaServer submitter is invalid.
- ***Server in maintenance***
- ...

To export the SpaServer synchronization results to notepad use the <***Export to Notepad***> button.

The *SpaServer synchronization* dialog box can be closed with the <***Close***> button.

8.4 Synchronizing with SpaServer (entry mode)

4.1 Double-click on a database entry to open the *Entry* window (see Figure 8.1). Right-clicking on the entry, and selecting ***Open entry*** also works.

4.2 In the *Entry* window select ***Spa-Typing*** > ***Synchronize with SpaServer***.

When no users are defined in the database, the error message 'No SpaServer users defined yet' is generated. In that event, select ***Spa-Typing*** > ***SpaServer synchronization settings*** in the *Main* window, and press the <***Add***> button to add a spa user to the database (see 8.2).

Prior to the submission of the data to the SpaServer, BIONUMERICS checks the presence of **BIONUMERICS errors** for this entry (see 8.3 for all possible BIONUMERICS errors):

- As a first check, BIONUMERICS checks if the experiment *Spa-typing* is present for the entry.
- In a second step, BIONUMERICS checks if the Spa-typing experiment contains an assembly.

If BIONUMERICS detects the presence of an assembly, the program checks if problems are present in the assembly and reports this:

- The assembly quality settings differ from the default settings that are required for the submission of new types to the SpaServer.
- The number of active traces in the contig is not equal to 2.
- One or both trimming patterns are not detected on the consensus sequence.
- The status of the Assembly reports an error (= red status box).

If none of the above described errors are present for the entry, BIONUMERICS checks in a next step if the information fields that are linked to one of the mandatory SpaServer information fields, contain (the correct) information.

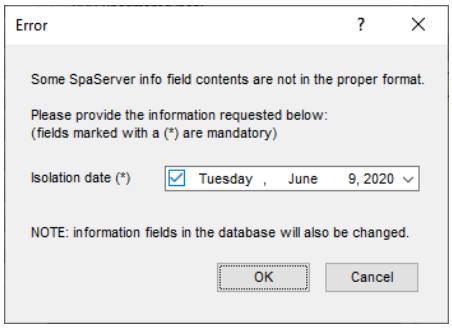


Figure 8.12: The *Error* dialog box.

When BIONUMERICS detects one or more information fields that do not contain the correct information, the *Error* dialog box pops ups, listing all the information fields that are not in the proper format. Select the correct information from the drop-down list(s) and press <***OK***>.

4.3 If no BIONUMERICS errors are detected, and if all fields are properly filled in, the *Submit to Ridom/SeqNet SpaServer* dialog box pops up (see Figure 8.13).

The first time spa information for an entry is submitted to the SpaServer, the *Submit to Ridom/SeqNet SpaServer* dialog box pops up, listing all spa information that BIONUMERICS will try to

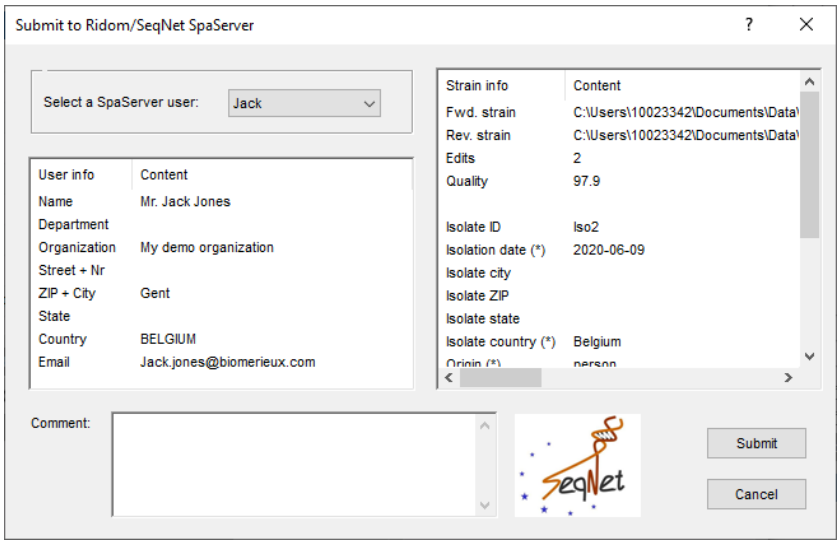


Figure 8.13: The *Submit to Ridom/SeqNet SpaServer* dialog box.

submit to the SpaServer (see Figure 8.13). The SpaServer submitter can be selected from the drop-down list in the left upper panel. Optionally, a comment can be entered in the ***Comment*** box.

The ***Quality*** score displayed in the list in the right panel is calculated based on the ***Assembly Quality*** settings. The quality of a strain is the percentage of bases in the consensus that has an average sequence base quality greater than or equal to 100. Spa data can only be submitted to the online SpaServer if the quality of the strain is greater than 70.

Pressing the <***Submit***> button submits the spa data to the online SpaServer. To cancel the submission, press the <***Cancel***> button.

4.4 Press the <***Submit***> button.

BIONUMERICS submits the spa data to the online SpaServer. Depending on the information that is submitted, **SpaServer response** errors may pop up (see 8.3 for all possible **SpaServer response** errors).

If (corrected) spa information is **re-submitted** to the SpaServer (e.g. when the server returned an error that was subsequently corrected), the *Submitted SpaServer data* dialog box pops up, displaying all previously submitted information. The ***SpaServer response*** of the submission is displayed in the ***Response*** field.

Pressing the <***Submit again***> button will pop up the *Submit to Ridom/SeqNet SpaServer* dialog box, listing all spa information for the new submission.

Bibliography

[1] G. Benson. Sequence alignment with tandem duplication. *Journal of Computational Biology*, 4(3):351–367, 1997.

bioMérieux S.A. • 69280 Marcy l'Étoile • France • Tel.: + 33 (0)4 78 87 20 00 • Fax: +33 (0)4 78 87 20 90
www.biomerieux.com