BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

Polymorphic VNTR typing plugin

Contents

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 *Polymorphic VNTR typing plugin*. Use of the plugin is supported in the **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE** configurations.

The features of the plugin will be illustrated using an example data set, which can be found on the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "TRST sample data files"). The data set was obtained from the Dublin Dental School and Hospital, Dublin, Ireland. The sequence files originate from *Staphylococcus aureus* and represent the mec-associated Direct Repeat Unit (DRU) region. The text files containing dru-repeat and dru-type information were obtained by courtesy of Prof. Goering, the Creighton University Medical Center, Omaha, NE, USA (http://www.dru-typing.org).

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***>.

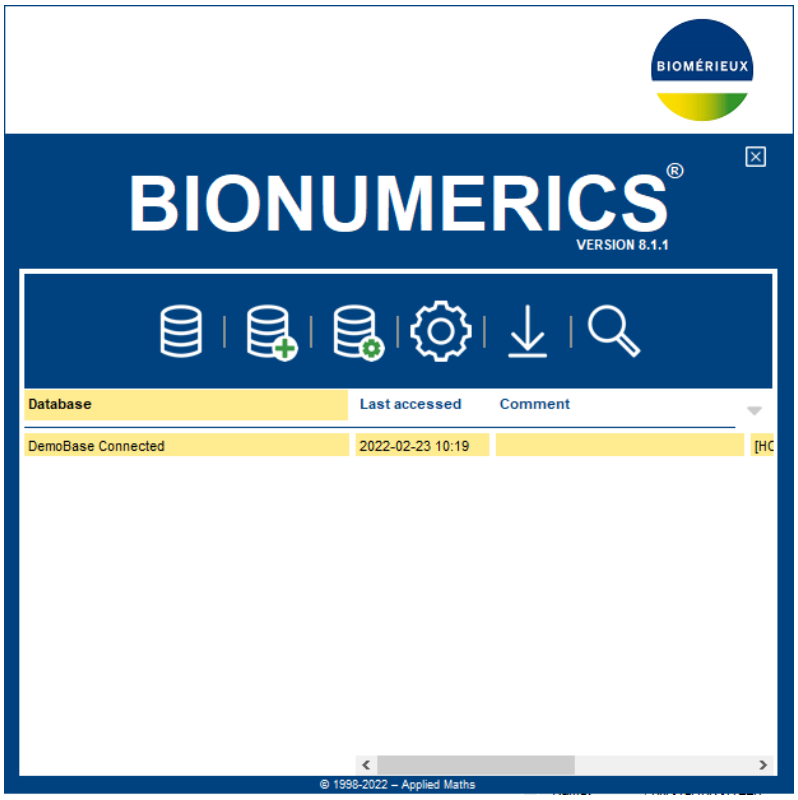


Figure 1.1: The *BIONUMERICS Startup* window.

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***>.

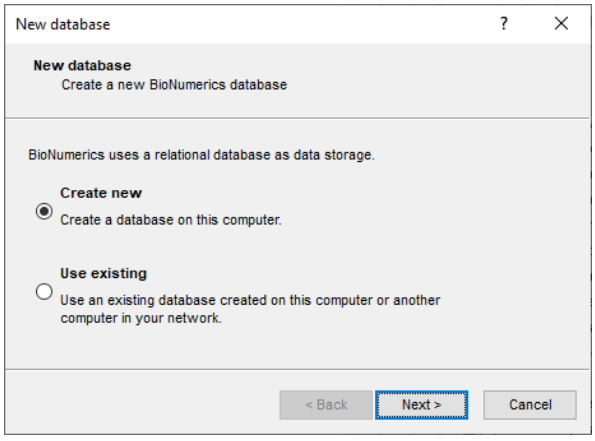


Figure 1.2: The *New database* wizard page.

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.

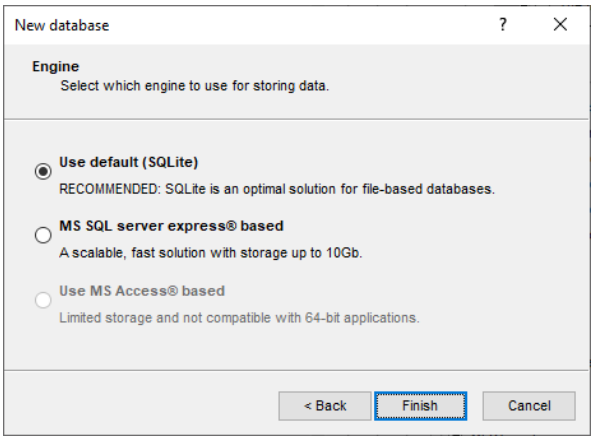


Figure 1.3: The *Engine* wizard page.

1.4 Installing the Polymorphic VNTR 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).

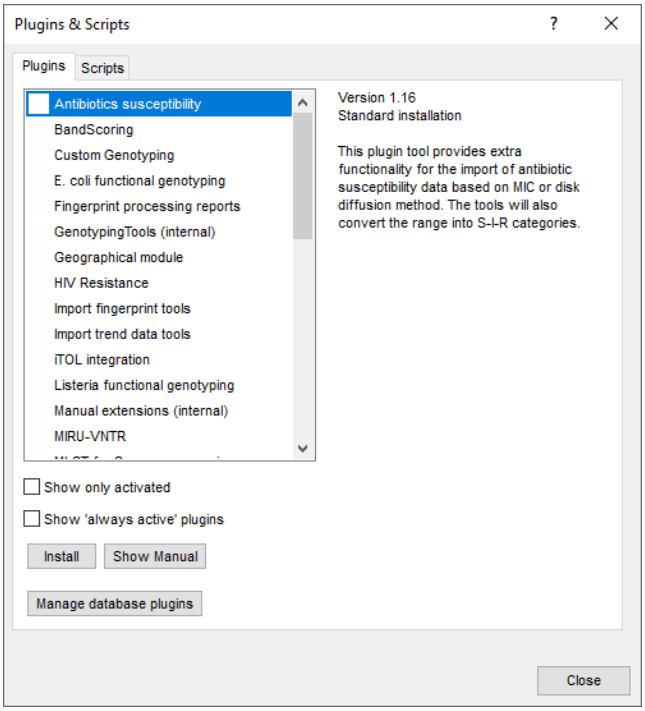


Figure 1.4: The *Plugins and Scripts* dialog box.

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 *Polymorphic VNTR 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.

4.3 When the *Polymorphic VNTR typing plugin* is successfully installed, a confirmation message pops up. Press <***OK***>.

4.4 Press <***Close***> to close the *Plugins and Scripts* dialog box and to continue to the *Main* window.

4.5 Close and reopen the database to activate the features of the *Polymorphic VNTR typing plugin*.

The *Polymorphic VNTR typing plugin* installs menu items in the main menu of the software under ***Repeat-Typing***.

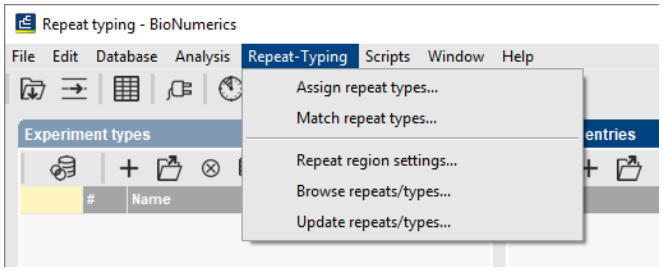


Figure 1.5: Repeat menu items in the *Main* window.

Chapter 2

Getting started

2.1 Repeat settings

1.1 Select ***Repeat-Typing*** > ***Repeat region settings*** to call the *Repeat regions* dialog box (see Figure 2.1).

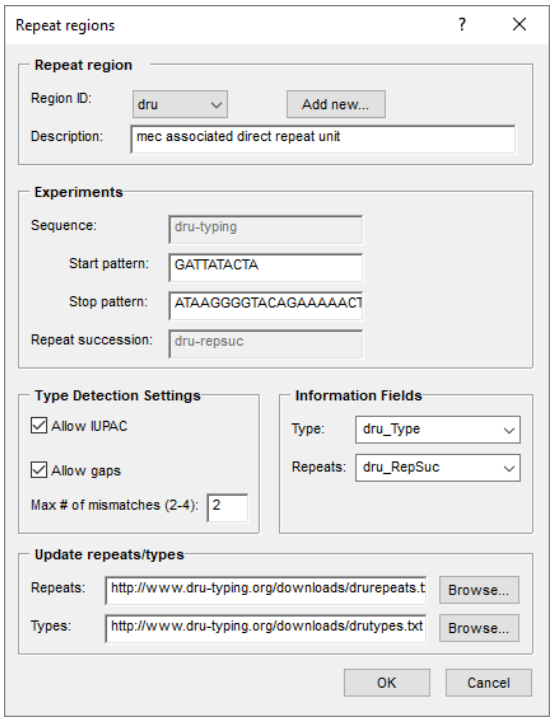


Figure 2.1: The *Repeat regions* dialog box.

Repeat region:

- In the upper part of the window, the name of the repeat region can be specified by pressing the <***Add new***> button next to ***Region ID***.
- Optionally a description can be specified in the ***Description*** text box.

Experiments:

- A sequence type is automatically created when a new region ID is specified in the *Repeat region panel*. The name of the sequence type is composed of the ***Region ID*** followed by the text ***-typing*** and is displayed in the ***Sequence*** text box. This sequence type will be used for the storage of the imported sequences.

- The trimming settings can be specified in the ***Start*** and ***Stop pattern*** boxes.

- An open character type is automatically created when a new region ID is specified in the *Repeat region panel*. The name of the character type is composed of the ***Region ID*** followed by the text ***-repsuc*** and is displayed in the ***Repeat succession*** text box. This character type will be used for the storage of the repeat successions.

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. 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***: A visualization tool is available in the plugin 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. Entered values higher than 4 will be set to 4.

Information Fields:

In this panel, you can choose the names of the database information fields that will hold the repeat type (***Type***), and the repeat succession string (***Repeats***) for your entries in the database. Choose the default suggested names, select an existing field, enter a new field name or set the options to ”None”.

The storage of a repeat succession in an information field is for illustration purposes only. 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 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 cannot start with a space.

Update repeats/types:

- Repeat and Type information available in text files or in online databases (e.g. Dru Server) can be uploaded to the BIONUMERICS database. The URL or file location of the ***Repeats*** and the ***Types*** can be specified in the *Update repeats/types panel*.

If you want to upload **repeat** information that is present in a text file, the repeat information needs to be present in FASTA format (see Figure 2.2). Avoid using ”-” signs in the repeat IDs and make sure no spaces or ”-” signs are present on the same line as the repeat sequences.

Figure 2.2: Dru repeats.

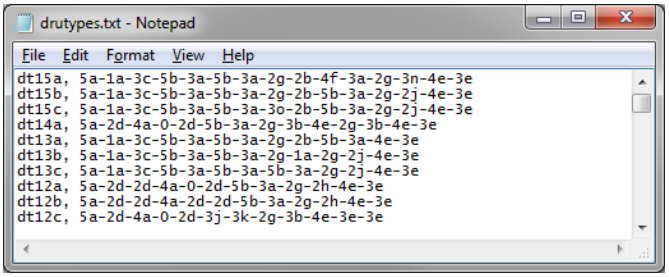


Figure 2.3: Dru types.

When uploading **type** information present in a text file, each line needs to correspond to a different type and each line needs to contain following information: **typeID**, **repeat-succession**, with the repeat succession being a string of ”-” separated repeat IDs (see Figure 2.3).

1.2 For this exercise, select the <***Add new***> button in the *Repeat region panel*.

This calls the *Add new repeat region* dialog box (see Figure 2.4).

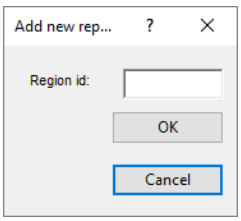


Figure 2.4: The *Add new repeat region* dialog box.

The *Add new repeat region* dialog box prompts for the repeat region name.

1.3 For this exercise enter the name ”dru” and press <***OK***>.

1.4 Optionally enter a ***Description*** in the *Repeat region panel*.

1.5 Enter the start pattern `GATTATACTA` and stop pattern `ATAAGGGGTACAGAAAAACT` in the *Experiments panel*.

1.6 For this exercise, enter the following URLs in the *Update repeats/types panel*:

- ***Repeats***: `http://www.dru-typing.org/downloads/drurepeats.txt`
- ***Types***: `http://www.dru-typing.org/downloads/drutypes.txt`

1.7 Leave all the settings unaltered in the other panels and press <***OK***>.

The information fields specified in the *Repeat region dialog box* are created and are displayed in the *Database entries* panel of the *Main* window. BIONUMERICS automatically creates a sequence type (for the import and storage of sequence data), and a character type (for the storage of the repeats). The experiments are listed in the *Experiment types* panel.

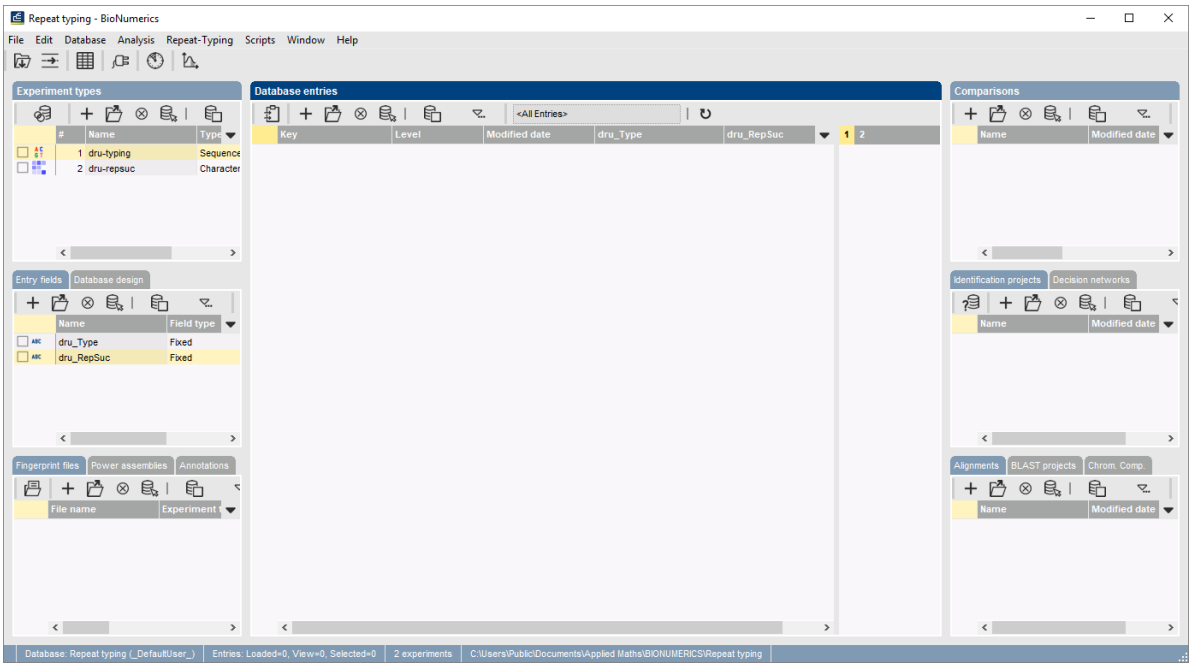


Figure 2.5: The *Main* window.

2.2 Update repeats and types

Repeat and Type information available in text files or in online databases (e.g. Dru Server) can be uploaded to the BIONUMERICS database.

2.1 Specify the URL or file location of the ***Repeats*** and/or the ***Types*** in the *Update repeats/types panel* of the *Repeat regions* dialog box (see Figure 2.1).

2.2 Select ***Repeat-Typing*** > ***Update repeats/types*** to update the repeats and/or types.

If more than one region is specified in the database, the *Select repeat region* dialog box is displayed (see Figure 2.6).

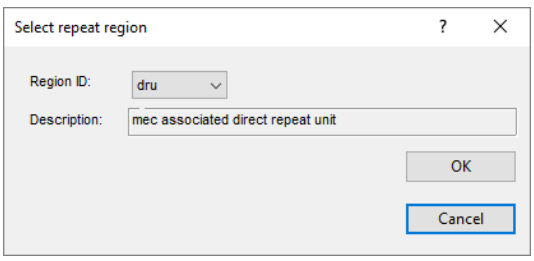


Figure 2.6: The *Select repeat region* dialog box.

In the *Select repeat region* dialog box all repeat regions that have been specified in the database are listed in the ***Region ID*** drop-down list. The ***Description*** of the selected region is displayed below.

2.3 If more than one repeat region has been specified, select the correct ***Region ID*** from the list and press <***OK***>.

If duplicate repeats or types are present in the text files or online databases the *Warning* dialog box appears (see Figure 2.7).

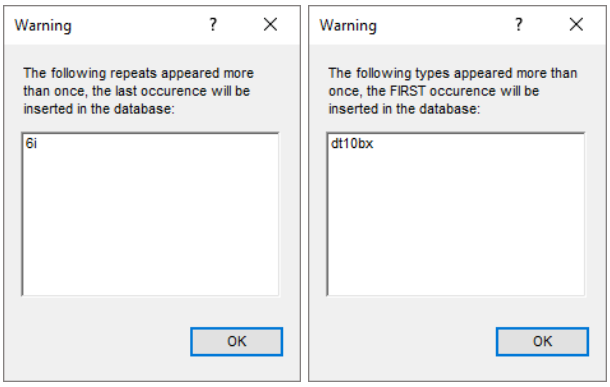


Figure 2.7: The *Warning* dialog box: repeats and types.

The *Warning* dialog box lists all duplicate repeats and types found in the selected text files or online databases. In case of duplicate repeats, only the last occurrence will be stored in the database; in case of duplicate types, only the first occurrence will be stored in the database.

2.4 If duplicate types/repeats are detected, press <***OK***>.

The repeat and type lists are updated. A confirmation message pops up.

2.5 Press <***OK***> once more.

2.3 Browse repeats or types

The lists of repeats and repeat types, uploaded from a file or online database, can be queried by the user.

3.1 Select ***Repeat-Typing*** > ***Browse repeats/types*** in the *Main* window. This action calls the *Browse types/repeats* dialog box (see Figure 2.8).

The ***Region ID*** should be selected from the drop-down list in the *Repeat region panel*.

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 uploaded to the database.

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

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

Edit the repeat information and press <***OK***> to save the changes.

Edit the type information and press <***OK***> to save the changes.

Use the <***Delete***> button to delete a repeat/type. Pressing the <***Delete all***> button deletes all repeats/types from the list.

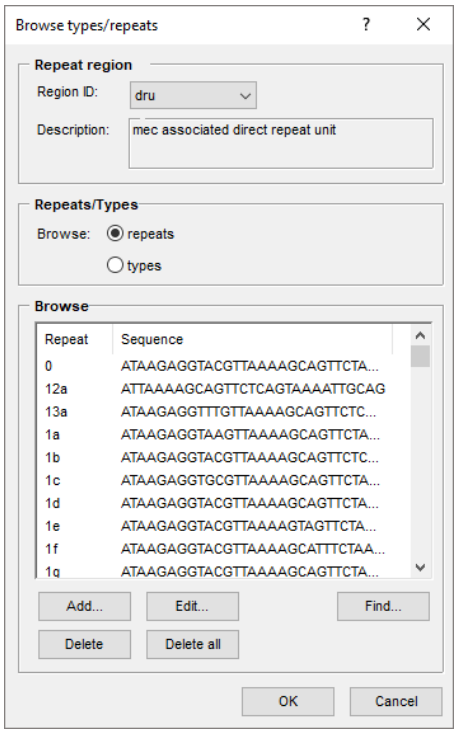


Figure 2.8: The *Browse types/repeats* dialog box.

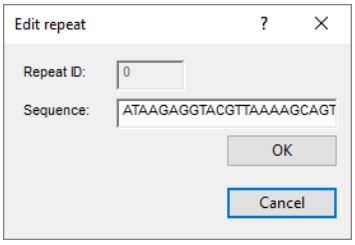


Figure 2.9: The *Edit repeat* dialog box.

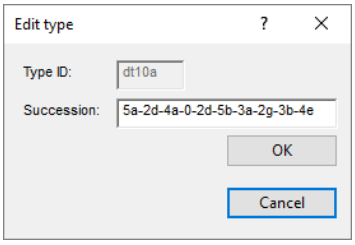


Figure 2.10: The *Edit type* dialog box.

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

Specify the repeat ID and sequence. Pressing <***OK***> saves the sequence to the database.

Specify the type ID and repeat succession. Pressing <***OK***> saves the type to the database.

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 ***Sequence*** text box and press the <***Find***> button. If the repeat is present in the list of repeats, the ***Repeat ID*** of the sequence is displayed.

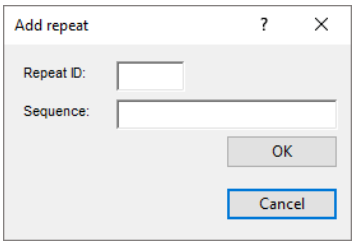


Figure 2.11: The *Add repeat* dialog box.

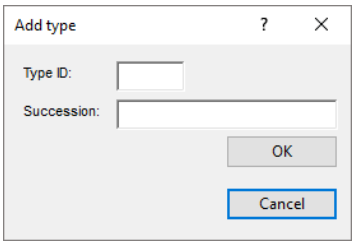


Figure 2.12: The *Add type* dialog box.

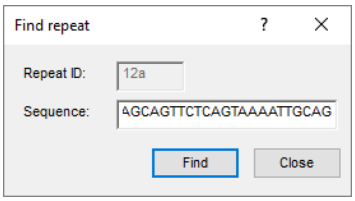


Figure 2.13: The *Find repeat* dialog box.

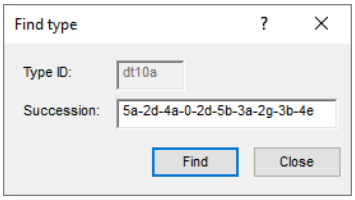


Figure 2.14: The *Find type* dialog box.

When looking for a type, enter the succession string in the ***Succession*** text 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.

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

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

3.3 As ***Object to report***, select ”TRST:Repeat sequences” or ”TRST: Sequence types” and press <***OK***> (see Figure 2.15).

For more information on object queries, see the Reference manual, Chapter Database objects.

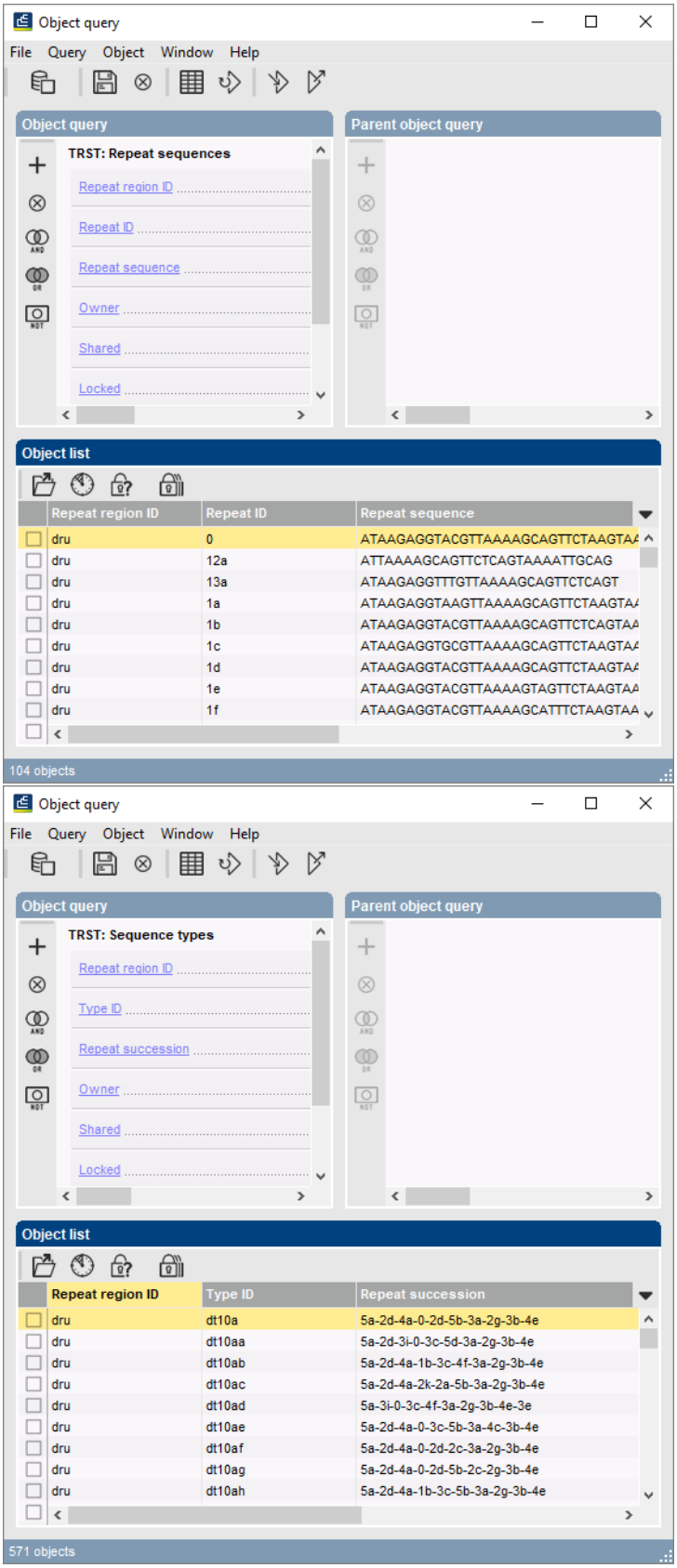


Figure 2.15: Object queries: Repeats and Types.

Chapter 3

Importing and assembling trace files

3.1 Importing and assembling trace files in batch

A set of trace files can be downloaded from the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "TRST sample data files") and are used in this guide to explain the work flow of the *Polymorphic VNTR 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 *Import sequence traces* dialog box is updated (see Figure 3.1).

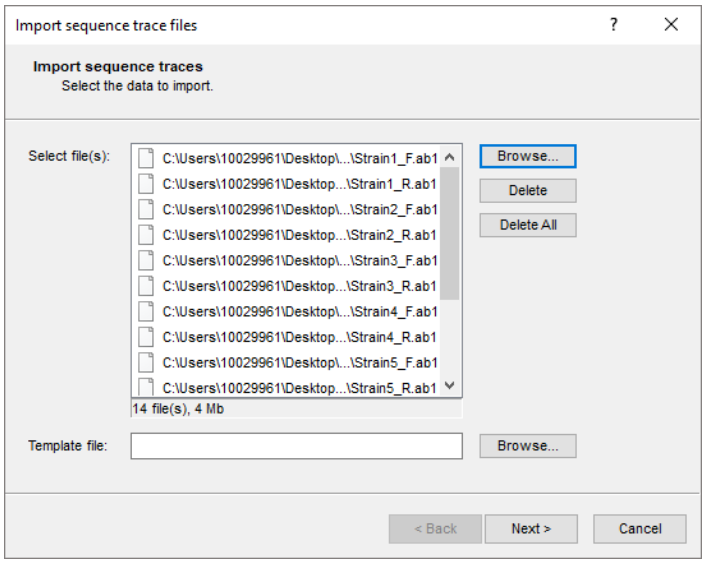


Figure 3.1: Select trace files.

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. A new import

template needs to be defined:

1.5 Press the ***Create new*** button to call the *Import rules* dialog box.

The only source of information available in the newly created import template is the file name.

1.6 Double-click on the ***Name*** row or select the row and press <***Edit destination***>. Select ***Key*** as destination and press <***OK***> (see Figure 3.2).

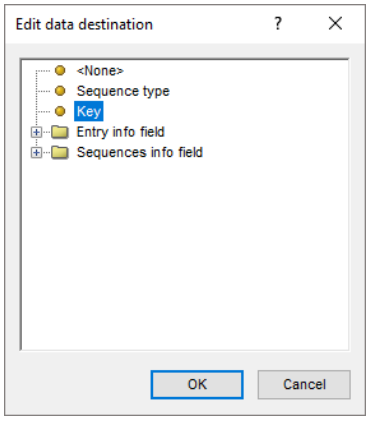


Figure 3.2: The *Edit data destination* dialog box.

The import rule in the *Import rules* dialog box is updated.

1.7 Check the option ***Show advanced options*** and press the <***Edit parsing***> button.

1.8 In the *Data parsing* dialog box, fill in following data parsing string: "[DATA]_*".

This parsing string will only take into account the text occurring before the first underscore (_). The asterisk (*) serves as a wildcard, meaning that all characters after the first underscore will be ignored.

1.9 Press the <***Preview***> button and press <***OK***> when the parsing is correct (see Figure 3.3).

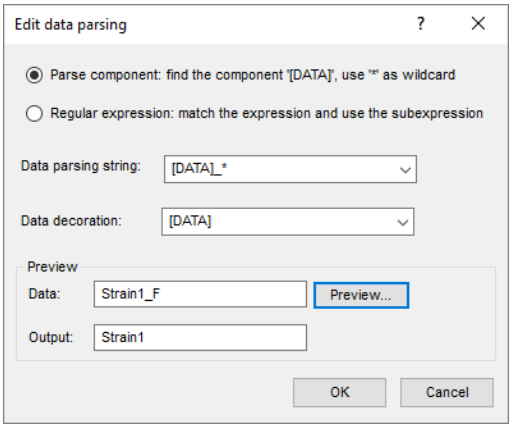


Figure 3.3: Data parsing string.

The *Import rules* dialog box should now look like Figure 3.4.

1.10 Press <***Next***> and <***Finish***>.

1.11 Specify a template name, e.g. **Import dru trace files** and press <***OK***>.

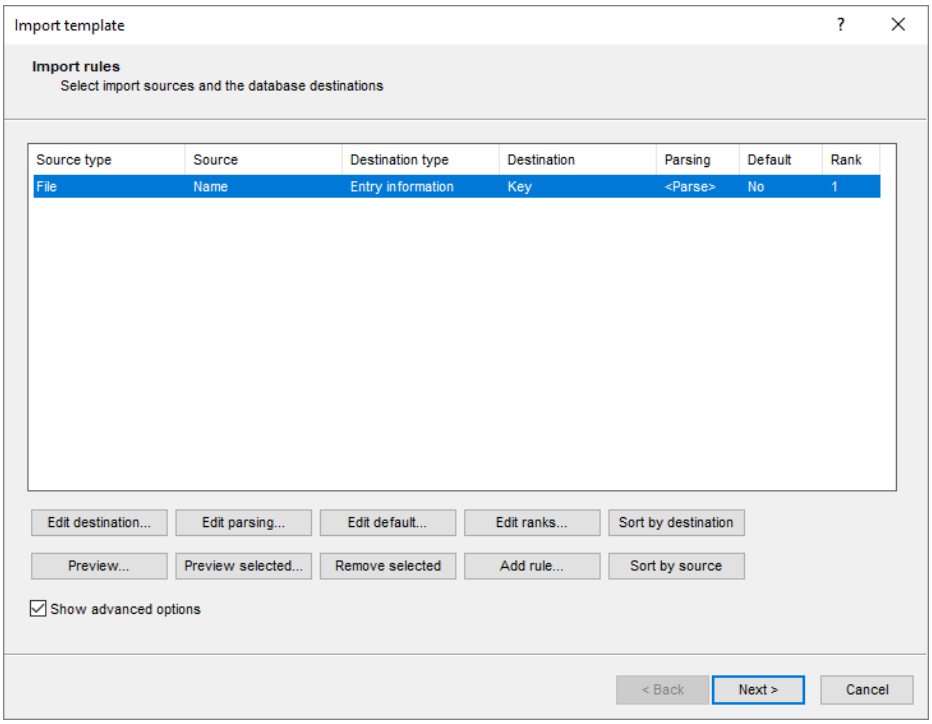


Figure 3.4: Import rules.

1.12 Make sure the newly created template is selected and press the <***Preview***> button.

The preview should now look like Figure 3.5.

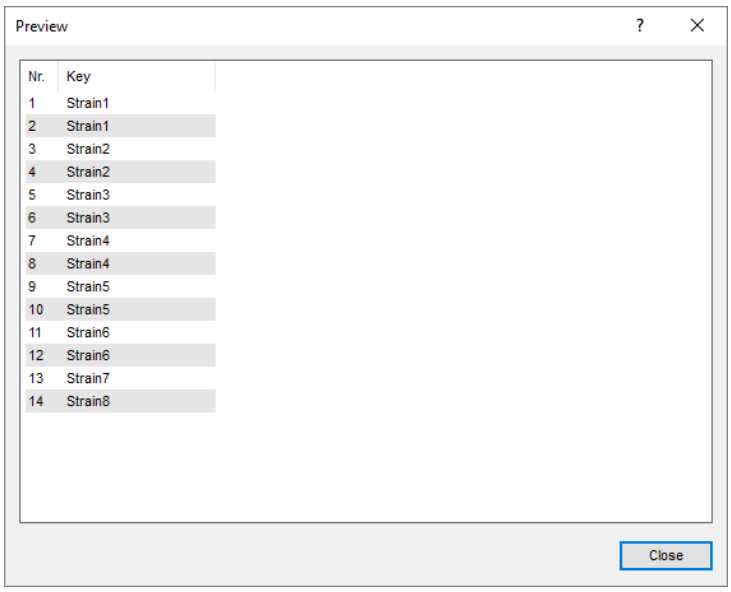


Figure 3.5: Preview of the parsing.

1.13 Close the preview.

1.14 Make sure the newly created template is selected, and select the ***dru-typing*** from the ***Experiment type*** list (see Figure 3.6).

1.15 Press <***Next***>.

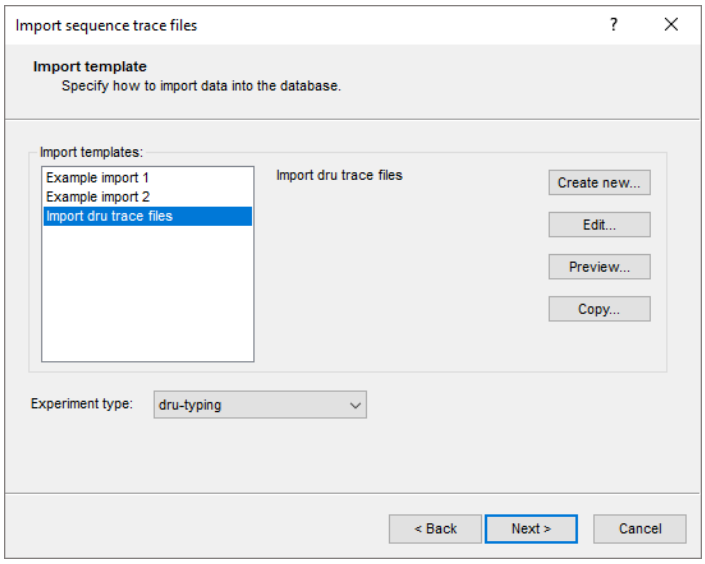


Figure 3.6: Import template.

1.16 Press <***Next***> once more to confirm the creation of 8 new entries (see Figure 3.7).

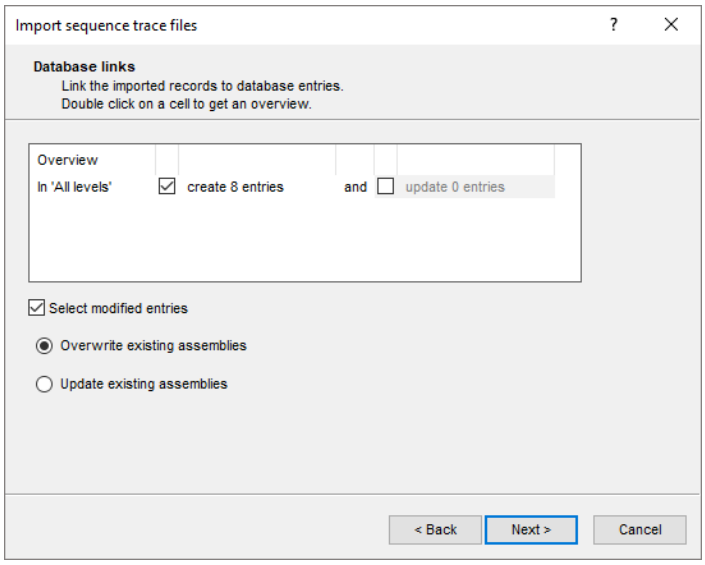


Figure 3.7: Database links.

The *Processing* dialog box opens (see Figure 3.8).

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 more sequences are different from the consensus.

1.17 Press <***Trimming settings***> to pop up the *Assembly trimming settings* dialog box (see Figure 3.9).

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

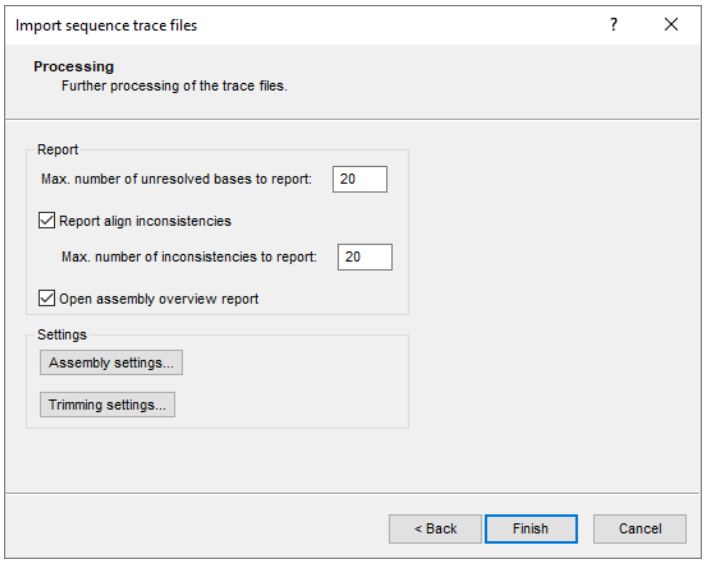


Figure 3.8: The *Processing* dialog box.

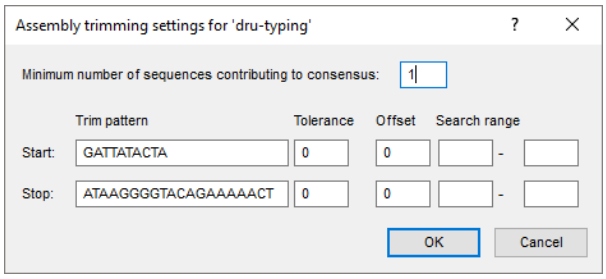


Figure 3.9: The *Assembly trimming settings* dialog box.

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 entered in the *Repeat regions* dialog box for the sequence type ***dru-typing*** (see Figure 2.1) are shown in the ***Start pattern*** and ***Stop pattern*** text boxes (see Figure 3.9).

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

1.19 Press the <***Assembly settings***> button to call the *Assembly settings* dialog box (see Figure 3.10).

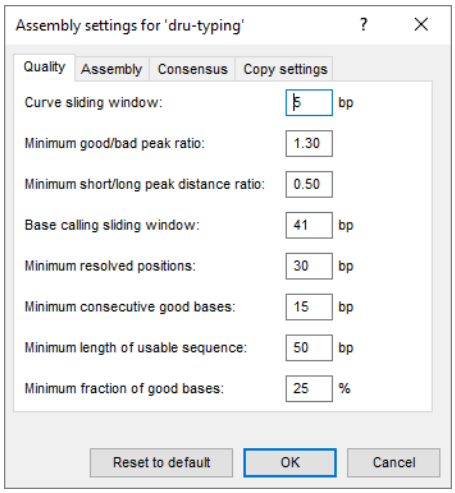


Figure 3.10: The *Assembly settings* dialog box.

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, see the Reference manual, Chapter Setting up sequence type experiments. In the last tab the Assembly settings can be copied from or to another sequence type experiment.

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

1.21 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 3.11). This window can also be displayed from the *Main* window with ***Analysis*** > ***Sequence types*** > ***Batch assembly reports...***.

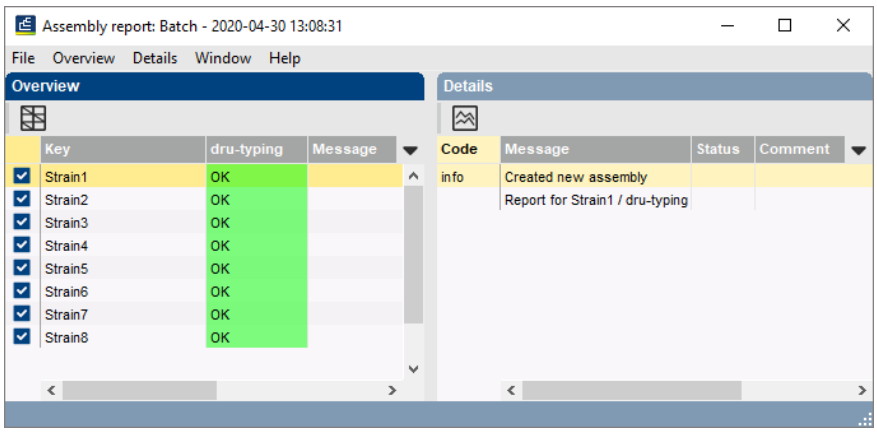


Figure 3.11: The *Batch sequence assembly report window* window.

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.
- **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. ***Strain1/dru-typing*** to update the *Details* panel on the right-hand side (see Figure 3.11).

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 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 **Strain1** 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.

4.2 Showing repeats on the consensus

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

Assembler screens the consensus sequence for repeats.

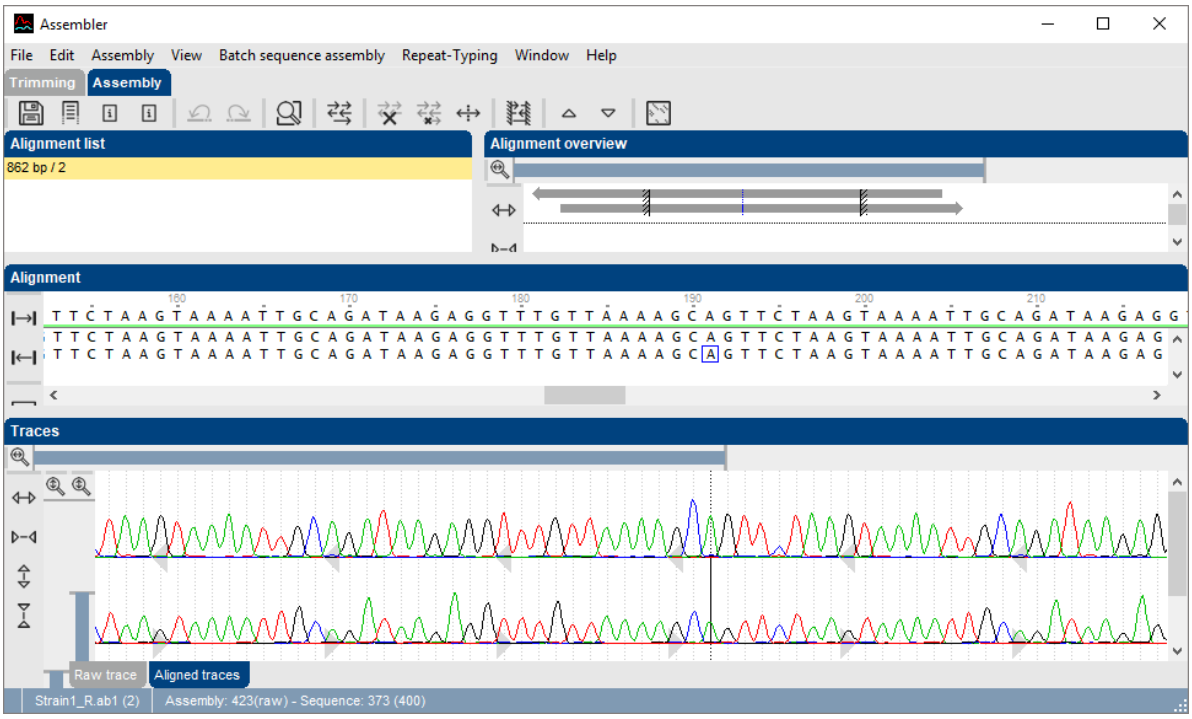


Figure 4.1: The *Aligned traces* panel.

- Known repeats are shown in *green* and the name of the repeat is shown on top of the know 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*.

Figure 4.2: Showing the repeats on the consensus sequence.

If the option ***Allow IUPAC code*** is checked in the *Repeat regions* dialog box (see Figure 2.1) 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.

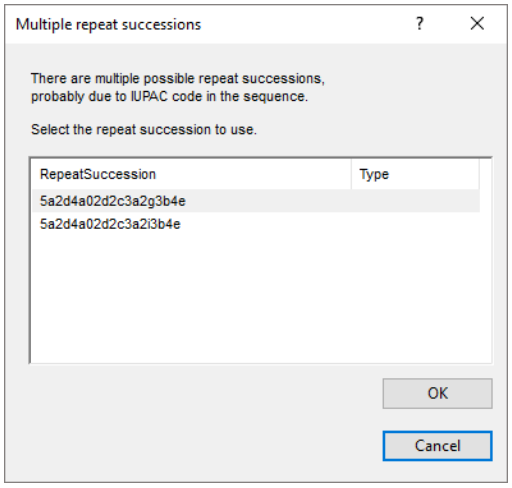


Figure 4.3: The *Multiple repeat successions* dialog box.

If the option ***Allow IUPAC code*** is checked in the *Repeat regions* dialog box (see Figure 2.1) 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?).

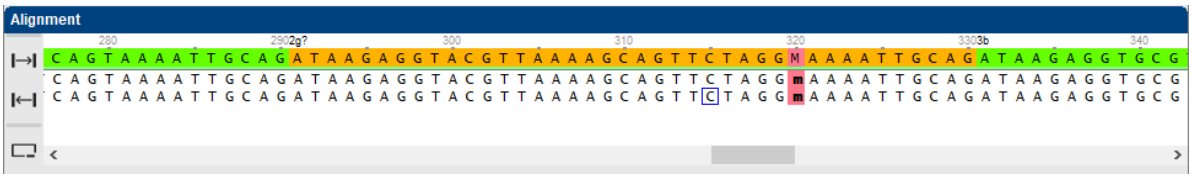


Figure 4.4: IUPAC code resulting in more than one known repeats.

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

Figure 4.5: Repeat succession string and repeat type.

When importing and assembling sequences, BIONUMERICS uses the parameters defined in the *Assembly settings* dialog box (see Figure 3.10).

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, Chapter Setting up sequence type experiments.

4.3 Showing the repeat succession plot

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

The repeats are displayed in the *Repeat plot window* (see Figure 4.6 and Figure 4.7).

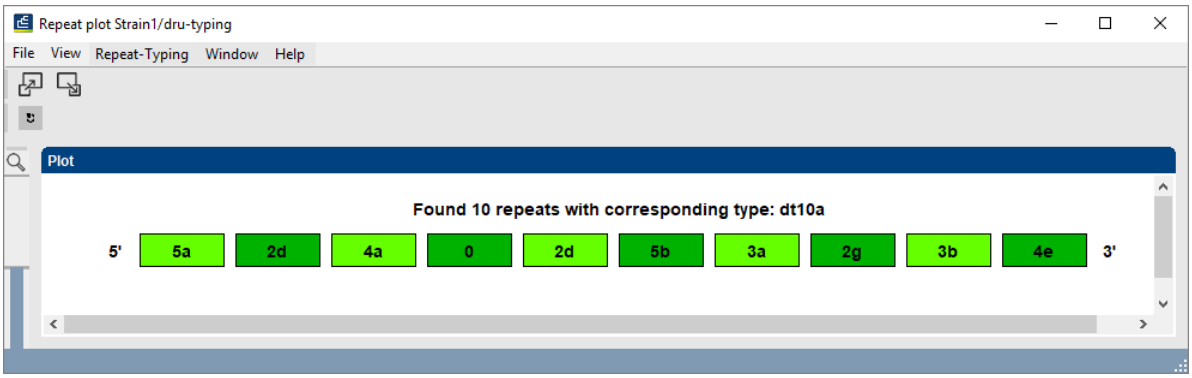
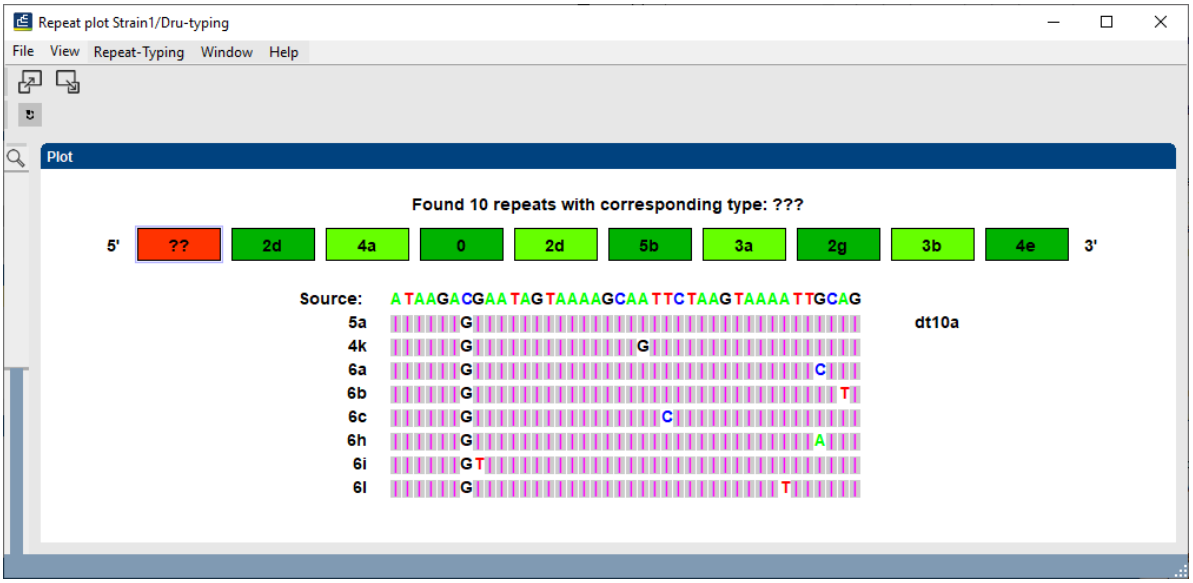


Figure 4.6: The repeat plot: 10 known repeats, corresponding to type dt10a.



(b)

Figure 4.7: The repeat plot: editing suggestions are displayed for the unknown repeat.

When clicking on an unknown red ”r??” repeat, a table is displayed with suggestions to edit the sequence (see Figure 4.7). In the left column, the repeat is shown. In the right column, the

associated repeat type - if available - is displayed.

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

Replacing the ”T” with an ”A” in the unknown repeat in Figure 4.7 results in repeat **5a** and repeat type **dt10a**. Looking at this position in the *Assembly view* gives additional information about the missing base.

3.3 When the consensus sequence has been edited in the *Contig assembly* window, select ***Repeat-Typing*** > ***Refresh*** in the repeat plot to update the repeat information.

More information on how to edit sequences in Assembler can be found in the Reference manual, Chapter Setting up sequence type experiments.

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

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

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

4.4 Changing the status of warning and error messages

4.4.1 Principles

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

- It is recommended to check the *warning* messages and solve them if needed. Since repeat 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 repeat 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 Option 1: Change 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 Option 2: Change status in the Detailed report window

4.2 After having solved all warnings and/or errors in Assembler, select ***File*** > ***Save*** (, **Ctrl+S**) and ***File*** > ***Exit*** to close the *Contig assembly* window.

4.3 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 4.8 for an example).

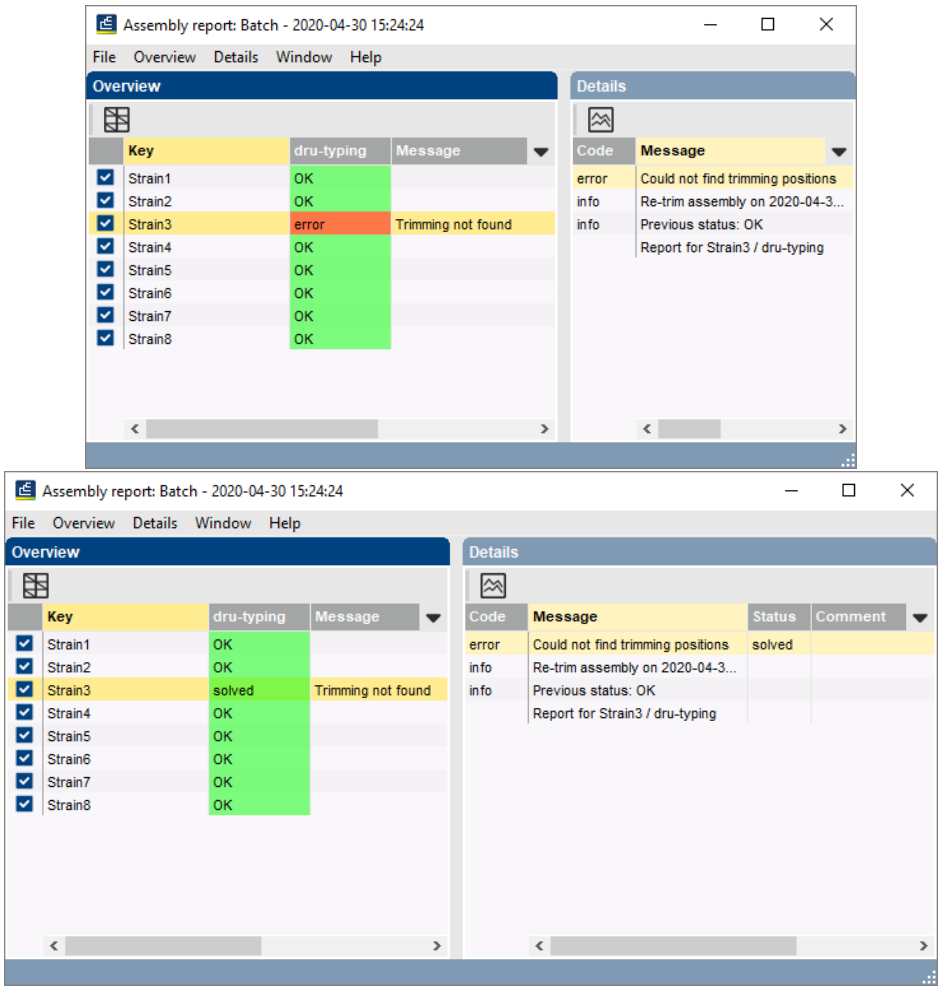


Figure 4.8: Solve errors/warnings.

Chapter 5

Repeat typing in BIONUMERICS

5.1 Selections in the main window

In the *Main* window, a repeat typing experiment (in our example: **dru-typing**) is present for each of the assembled sequences (see colored dot in the second column in the *Experiment presence* panel).

Screening for 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 types

5.2.1 Principles

2.1 Make a selection in the *Main* window.

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

The *Select repeat region* dialog box pops up (see Figure 5.1).

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.

In the *Select repeat region* dialog box all repeat regions that have been specified in the database are listed in the ***Region ID*** drop-down list. The ***Description*** of the selected region is displayed

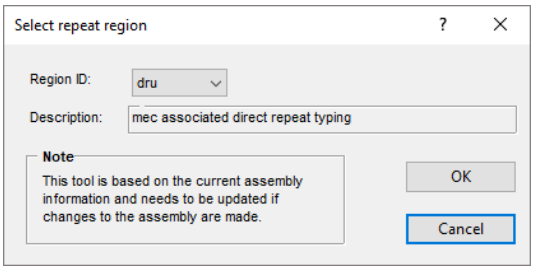


Figure 5.1: The *Select repeat region* dialog box.

below.

2.3 For this exercise, the ***dru*** region is the only region ID defined in the database and is automatically selected. Press <***OK***>.

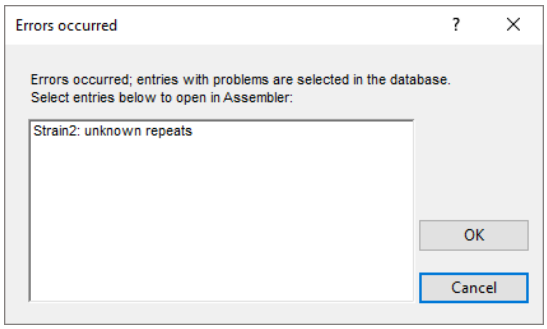


Figure 5.2: The *Errors occurred* dialog box.

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). Repeat 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 *Polymorphic VNTR typing plugin* uses a 2-step approach when the command ***Repeat-Typing*** > ***Assign types*** is selected:

5.2.2 Step 1: The assembly is screened for repeats

The repeat succession is stored in the character type ”repeatID-repsuc” (in this exercise: **dru-repsuc**) *and* the succession is displayed in the database information field that holds the repeat succession information (in this exercise: **dru_RepSuc**).

2.4 Click on the colored dot in the **dru_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.

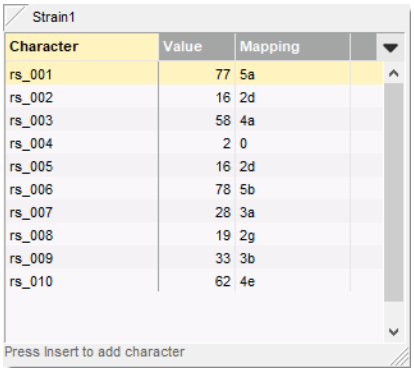


Figure 5.3: The **dru_RepSuc** character card, displaying the repeat succession in the ***Mapping*** column.

When a repeat does not match one of the repeats in the database, or when a IUPAC code is present in the consensus sequence, a "??" is placed at this position in the repeat succession information field and in the ***Mapping*** column of the character card.

When a sequence is found that is too short or too long to be considered as a repeat sequence, an asterisk (*) is placed at this position in the repeat succession information field and in the ***Mapping*** column of the character card.

When no repeats are found, no information is written in the repeat succession information field.

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

The repeat type is displayed in the information field that holds the repeats type information (in this exercise: **dru_Type**).

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

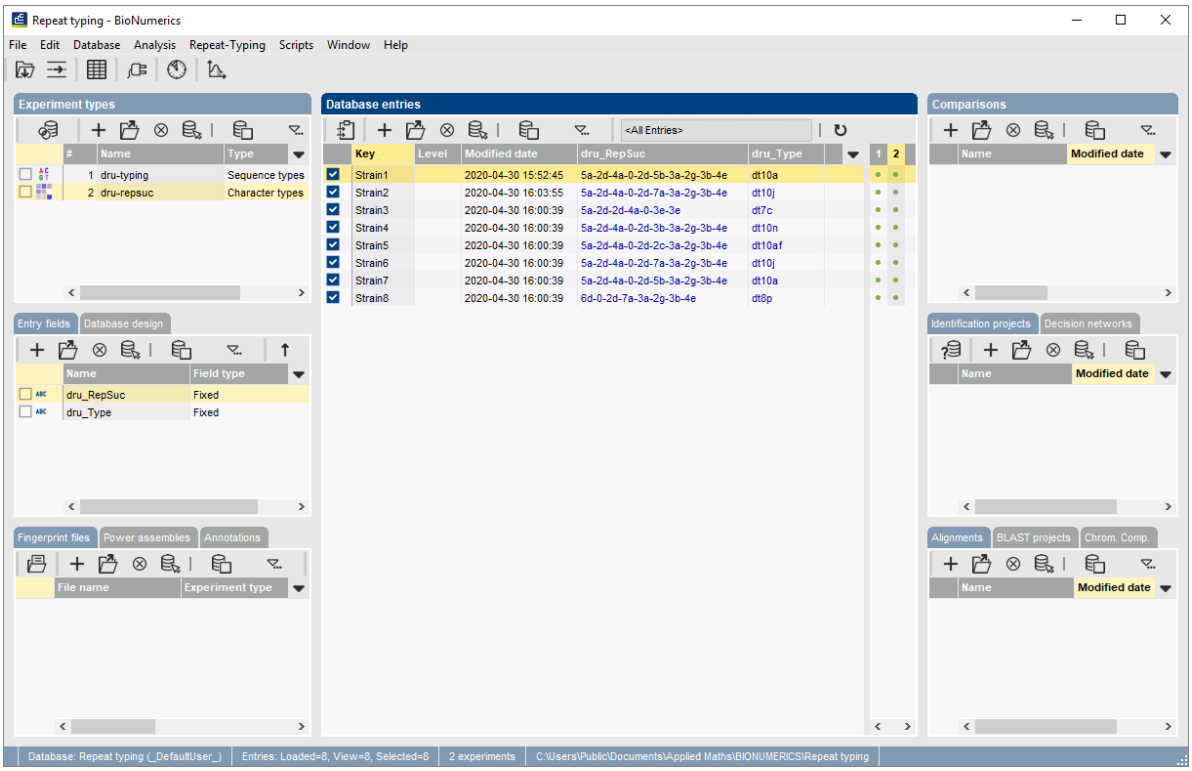


Figure 5.4: The *Main* window after assignment of repeats and repeat types.

Chapter 6

Cluster analysis of repeat types

6.1 Introduction

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

The *Polymorphic VNTR 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 repeats. This *DSI 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.

In our database, two experiment types are available and are shown in the *Experiments* panel.

2.5 Click on the eye button ( ) of the character type ”regionID-repsuc” (in this exercise: **dru-repsuc**).

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, Chapter Cluster analysis of characters 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).

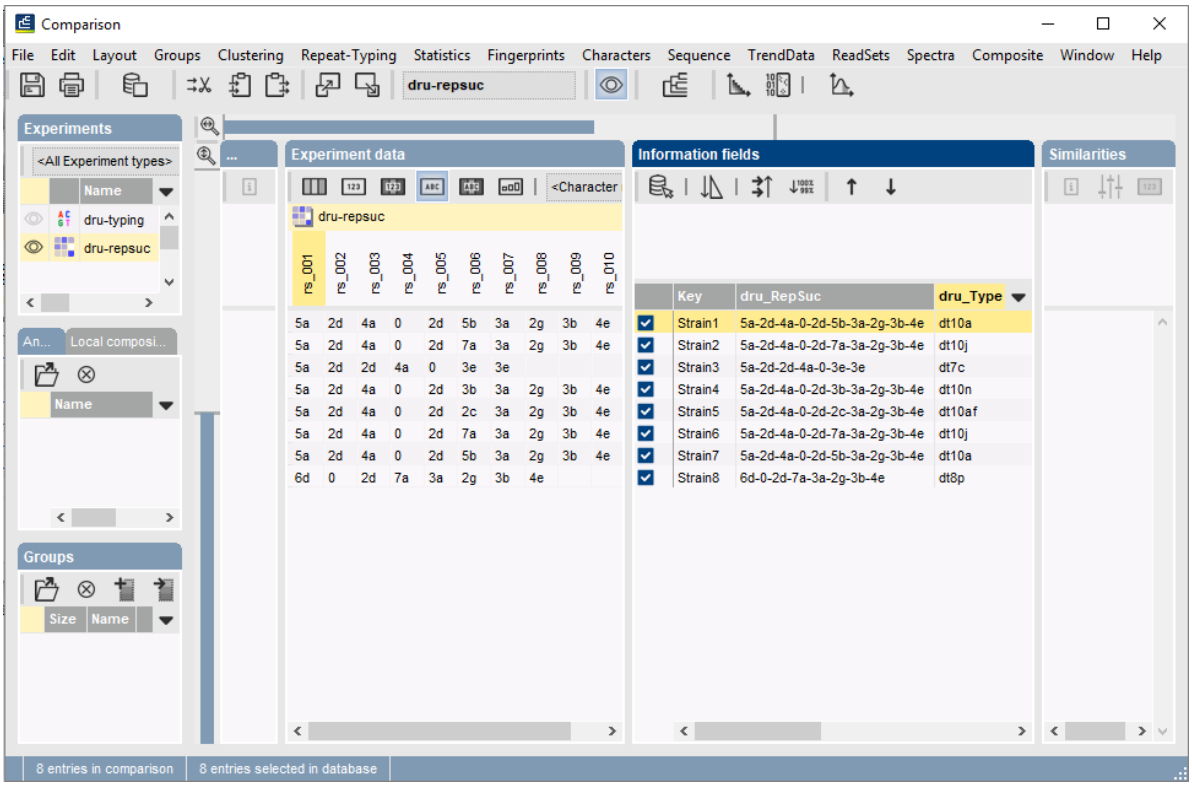


Figure 6.1: The *Comparison* window.

6.3 Creating a cost matrix

In the *Polymorphic VNTR 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 ***Repeat-Typing*** > ***Cost matrices*** in the *Comparison* window for the creation of your own cost matrix.

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

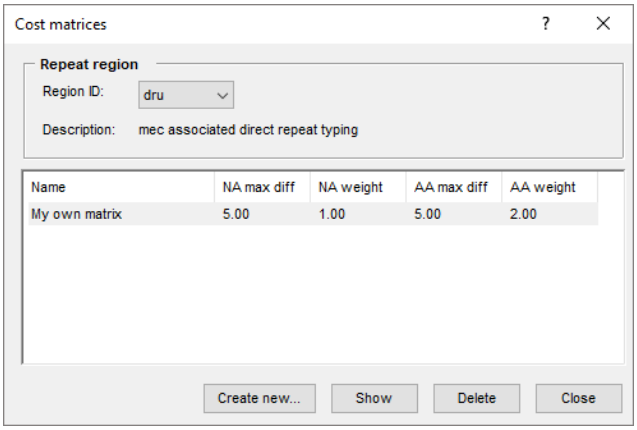


Figure 6.2: The *Cost matrices* dialog box.

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

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

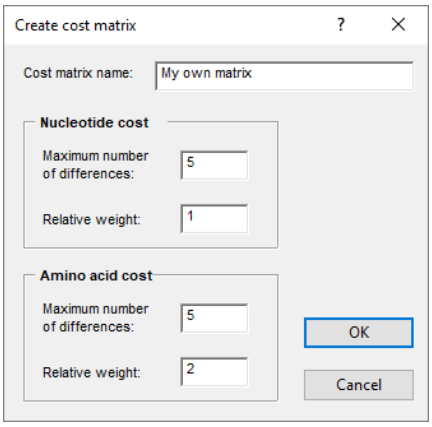


Figure 6.3: The *Create cost matrix* dialog box.

In the *Create cost matrix* dialog box 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 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).

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 (see Figure 6.5).

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

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

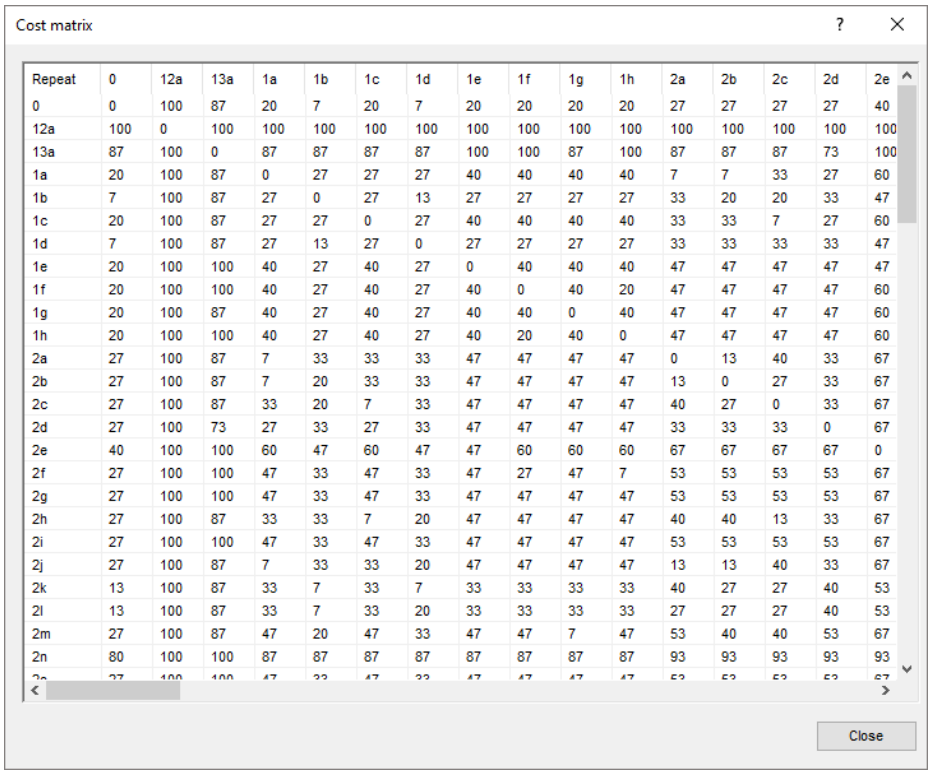
6.4 Cluster analysis settings

4.1 Select ***Repeat-Typing*** > ***Cluster types*** in the *Comparison* window.

If more than one region is specified in the database, the *Select repeat region* dialog box is displayed.

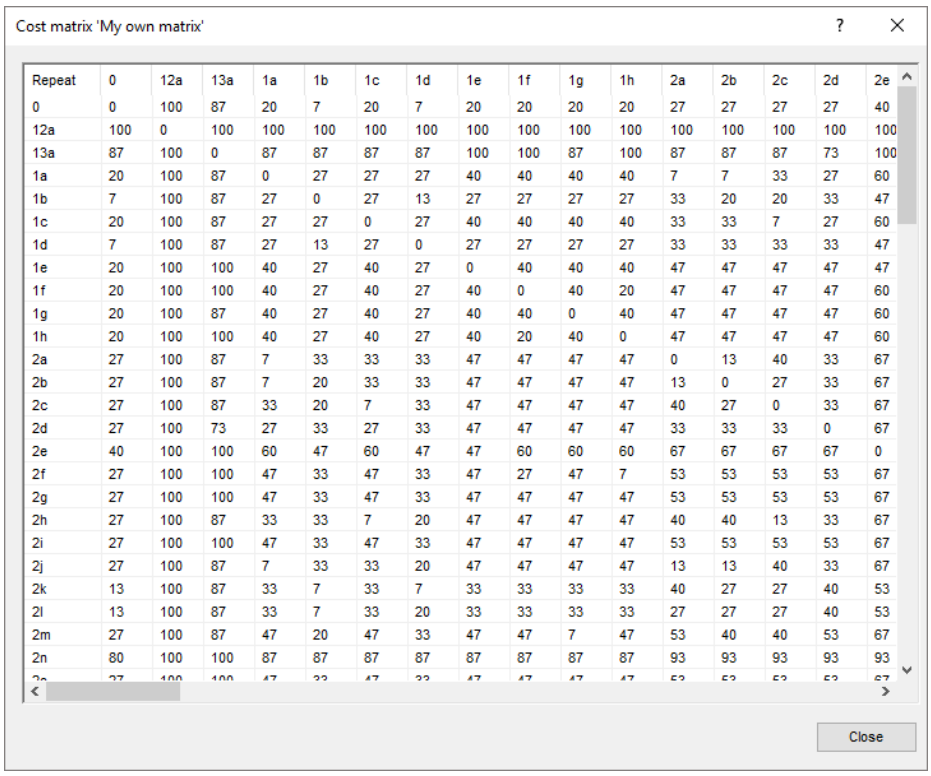
4.2 Select the correct ***Region ID*** from the list and press <***OK***>.

The *Clustering* dialog box appears (see Figure 6.6).



Repeat	0	12a	13a	1a	1b	1c	1d	1e	1f	1g	1h	2a	2b	2c	2d	2e
0	0	100	87	20	7	20	7	20	20	20	20	27	27	27	27	40
12a	100	0	100	100	100	100	100	100	100	100	100	100	100	100	100	100
13a	87	100	0	87	87	87	87	100	100	87	100	87	87	87	73	100
1a	20	100	87	0	27	27	27	40	40	40	40	7	7	33	27	60
1b	7	100	87	27	0	27	13	27	27	27	27	33	20	20	33	47
1c	20	100	87	27	27	0	27	40	40	40	40	33	33	7	27	60
1d	7	100	87	27	13	27	0	27	27	27	27	33	33	33	33	47
1e	20	100	100	40	27	40	27	0	40	40	40	47	47	47	47	47
1f	20	100	100	40	27	40	27	40	0	40	20	47	47	47	47	60
1g	20	100	87	40	27	40	27	40	40	0	40	47	47	47	47	60
1h	20	100	100	40	27	40	27	40	20	40	0	47	47	47	47	60
2a	27	100	87	7	33	33	33	47	47	47	47	0	13	40	33	67
2b	27	100	87	7	20	33	33	47	47	47	47	13	0	27	33	67
2c	27	100	87	33	20	7	33	47	47	47	47	40	27	0	33	67
2d	27	100	73	27	33	27	33	47	47	47	47	33	33	33	0	67
2e	40	100	100	60	47	60	47	47	60	60	60	67	67	67	67	0
2f	27	100	100	47	33	47	33	47	27	47	7	53	53	53	53	67
2g	27	100	100	47	33	47	33	47	47	47	47	53	53	53	53	67
2h	27	100	87	33	33	7	20	47	47	47	47	40	40	13	33	67
2i	27	100	100	47	33	47	33	47	47	47	47	53	53	53	53	67
2j	27	100	87	7	33	33	20	47	47	47	47	13	13	40	33	67
2k	13	100	87	33	7	33	7	33	33	33	33	40	27	27	40	53
2l	13	100	87	33	7	33	20	33	33	33	33	27	27	27	40	53
2m	27	100	87	47	20	47	33	47	47	7	47	53	40	40	53	67
2n	80	100	100	87	87	87	87	87	87	87	87	93	93	93	93	93

Figure 6.4: The *Cost matrix* dialog box.



Repeat	0	12a	13a	1a	1b	1c	1d	1e	1f	1g	1h	2a	2b	2c	2d	2e
0	0	100	87	20	7	20	7	20	20	20	20	27	27	27	27	40
12a	100	0	100	100	100	100	100	100	100	100	100	100	100	100	100	100
13a	87	100	0	87	87	87	87	100	100	87	100	87	87	87	73	100
1a	20	100	87	0	27	27	27	40	40	40	40	7	7	33	27	60
1b	7	100	87	27	0	27	13	27	27	27	27	33	20	20	33	47
1c	20	100	87	27	27	0	27	40	40	40	40	33	33	7	27	60
1d	7	100	87	27	13	27	0	27	27	27	27	33	33	33	33	47
1e	20	100	100	40	27	40	27	0	40	40	40	47	47	47	47	47
1f	20	100	100	40	27	40	27	40	0	40	20	47	47	47	47	60
1g	20	100	87	40	27	40	27	40	40	0	40	47	47	47	47	60
1h	20	100	100	40	27	40	27	40	20	40	0	47	47	47	47	60
2a	27	100	87	7	33	33	33	47	47	47	47	0	13	40	33	67
2b	27	100	87	7	20	33	33	47	47	47	47	13	0	27	33	67
2c	27	100	87	33	20	7	33	47	47	47	47	40	27	0	33	67
2d	27	100	73	27	33	27	33	47	47	47	47	33	33	33	0	67
2e	40	100	100	60	47	60	47	47	60	60	60	67	67	67	67	0
2f	27	100	100	47	33	47	33	47	27	47	7	53	53	53	53	67
2g	27	100	100	47	33	47	33	47	47	47	47	53	53	53	53	67
2h	27	100	87	33	33	7	20	47	47	47	47	40	40	13	33	67
2i	27	100	100	47	33	47	33	47	47	47	47	53	53	53	53	67
2j	27	100	87	7	33	33	20	47	47	47	47	13	13	40	33	67
2k	13	100	87	33	7	33	7	33	33	33	33	40	27	27	40	53
2l	13	100	87	33	7	33	20	33	33	33	33	27	27	27	40	53
2m	27	100	87	47	20	47	33	47	47	7	47	53	40	40	53	67
2n	80	100	100	87	87	87	87	87	87	87	87	93	93	93	93	93

Figure 6.5: The *Cost matrix* dialog box.

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

Alignment settings:

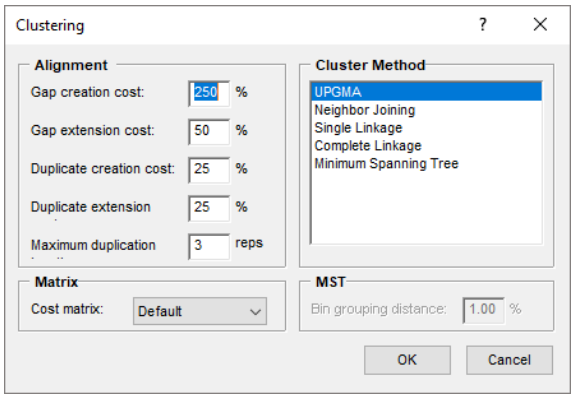


Figure 6.6: The *Clustering* dialog box.

- ***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 ***Repeat-Typing*** > ***Cluster types*** in the *Comparison* window.

If more than one region is specified in the database, the *Select repeat region* dialog box is displayed.

5.2 Select the correct ***Region ID*** from the list and press <***OK***>.

5.3 Select ***Minimum Spanning Tree*** in the *Cluster Method panel* (see 6.4).

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%.

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

The *Cluster analysis* window pops up (see Figure 6.7). 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).

5.5 Select a node or branch by clicking on them. To select several nodes/branches hold 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, Chapter The Advanced cluster analysis window.

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

5.7 In the *Node labels and sizes tab*, select ***Show node labels*** and select ***dru Type*** from the list.

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

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

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

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

The *Cluster analysis* window should now look like Figure 6.7.

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

5.13 Close 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 *Polymorphic VNTR 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 6.6).

6.1 In the *Comparison* window, choose ***Repeat-Typing*** > ***Cluster types***.

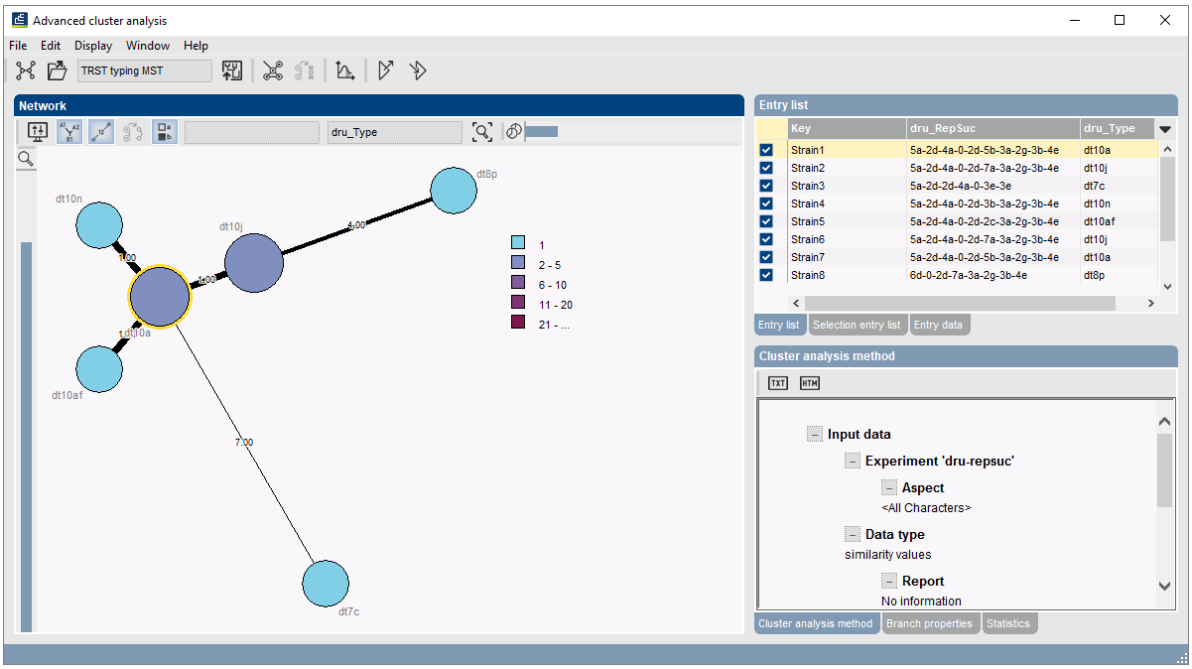


Figure 6.7: The *Cluster analysis* window.

If more than one region is specified in the database, the *Select repeat region* dialog box is displayed.

6.2 Select the correct ***Region ID*** from the list and press <***OK***>.

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

The dendrogram is shown in the *Comparison* window (see Figure 6.8).

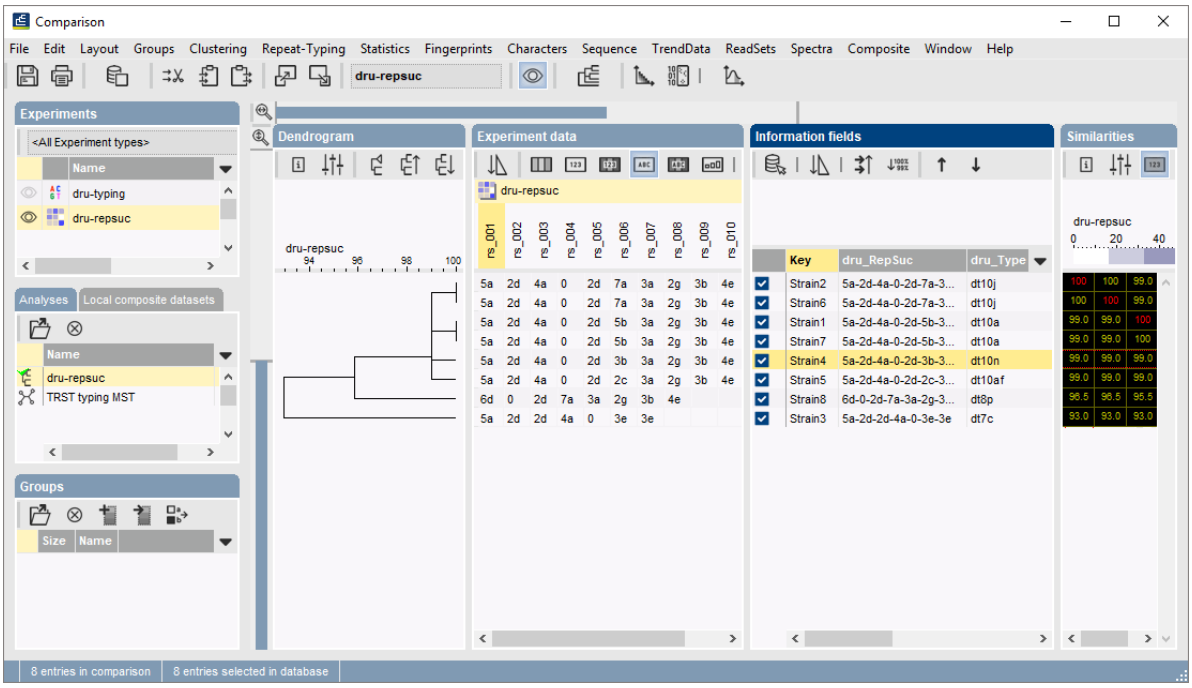


Figure 6.8: The *Comparison* window: dendrogram and similarity matrix.

6.4 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, Chapter Comparisons in BIONUMERICS.

6.5 Save and close the *Comparison* window.

Chapter 7

Matching repeat types

7.1 Selections in the main window

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

1.1 As an exercise, select a few strains in the *Main* window using the **Ctrl-** key (e.g. **Strain1**, **Strain4**, and **Strain5**).

7.2 Match types

2.1 Call the *Matching* dialog box with ***Repeat-Typing*** > ***Match repeat types***.

If more than one region is specified in the database, the *Select repeat region* dialog box is displayed.

2.2 Select the correct ***Region ID*** from the list and press <***OK***>.

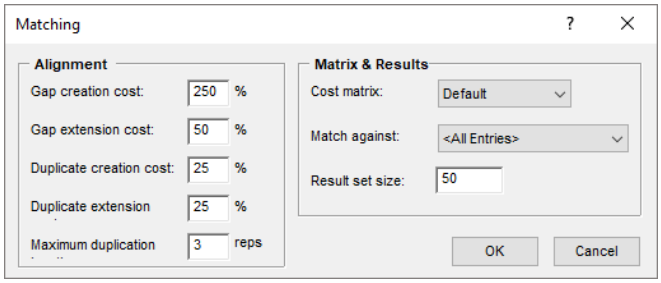


Figure 7.1: The *Matching* dialog box.

In the *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 types***>).
- ***Result set size***: Defines the number of best matches that are shown in the detailed report.

2.3 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 the repeats that are present in the associated character type experiment. The *Matching window* appears(see Figure 7.2).

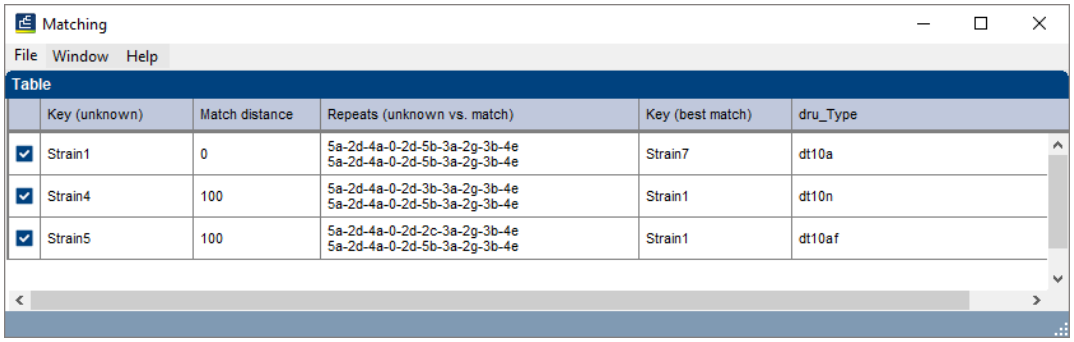


Figure 7.2: The *Matching window*.

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

Double-clicking an entry opens a detailed report (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.

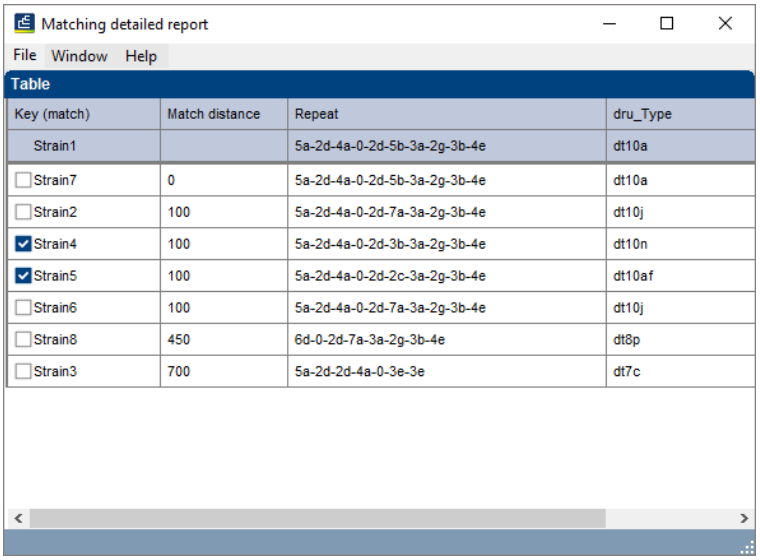


Figure 7.3: The *Detailed matching window*.

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