BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

Import fingerprint tools 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 *Import fingerprint tools plugin* of BIONUMERICS. This plugin allows you to install additional ***Fingerprint type*** import routines in the *Import data* wizard which can be accessed with ***File*** > ***Import...*** (, **Ctrl+I**).

The *Import fingerprint tools plugin* is supported in all configurations except the **BIONUMERICS-SEQ** configuration.

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 Installing the Import fingerprint tools plugin

Installing a plugin in a BIONUMERICS database is done from the *Plugins and Scripts* dialog box (see Figure 1.2), which can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ).

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.

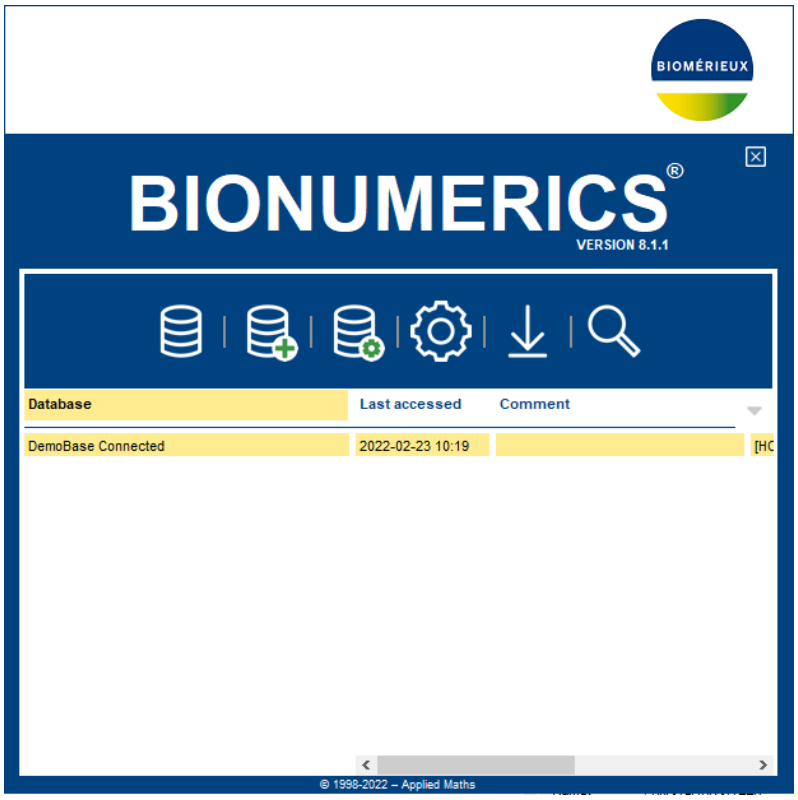

Figure 1.1: The *BIONUMERICS Startup* window.

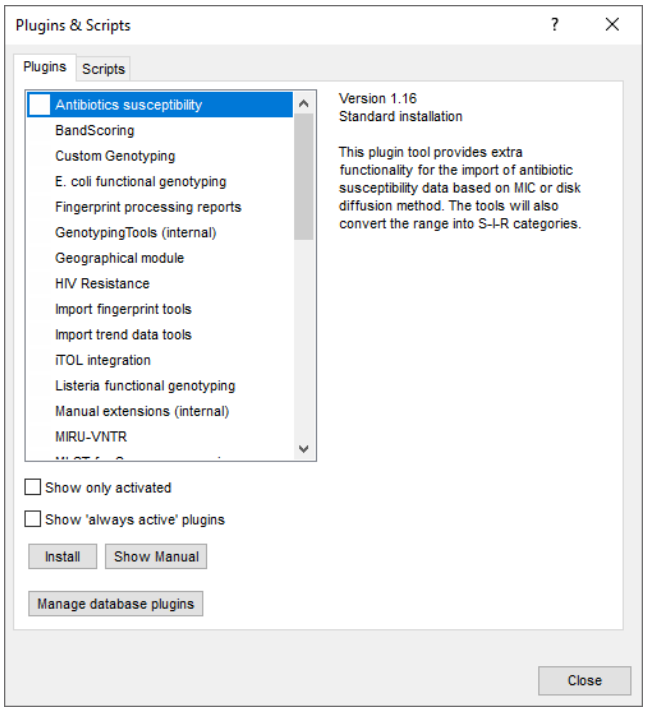

Figure 1.2: The *Plugins and Scripts* dialog box.

Proceed as follows to install the *Import fingerprint tools plugin*, starting from the *Plugins and Scripts* dialog box:

3.1 Select the *Import fingerprint tools plugin* in the list and press the <***Install***> button.

3.2 Confirm the installation of the plugin.

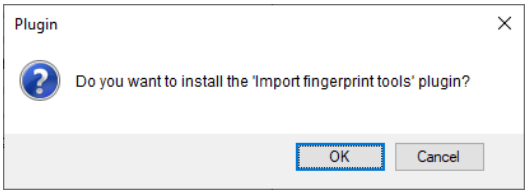

Figure 1.3: Confirm installation of the *Import fingerprint tools plugin*.

3.3 Confirm the installation of the plugin and press <***OK***>.

Once the plugin is successfully installed, it is marked with a green V-sign in the *Plugins and Scripts* dialog box (see Figure 1.4).

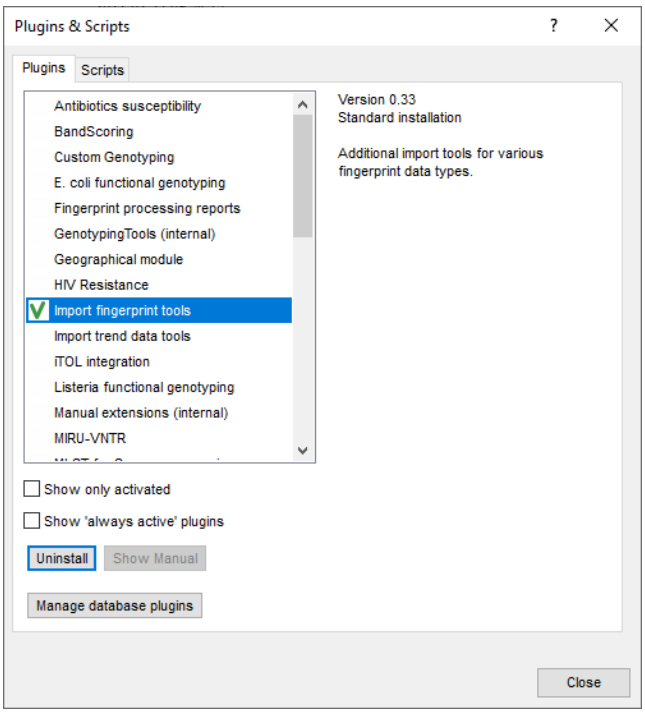

Figure 1.4: Installed plugin.

3.4 Close the *Plugins and Scripts* dialog box.

3.5 Close and reopen the database to activate the features of the *Import fingerprint tools plugin*.

Chapter 2

Import of AATI FragmentAnalyzer curve files

2.1 Introduction

The Fragment Analyzer is a capillary electrophoresis instrument from the company Advanced Analytical Technologies (http://www.aati-us.com). The generated electropherograms can be exported from the AATI PROSize® data analysis software in csv or txt format and imported in BIONUMERICS after installation of the *Import fingerprint tools plugin*.

2.2 Setting up a FragmentAnalyzer fingerprint type experiment

Next a new fingerprint type experiment will be created that will hold the imported Fragment Analyzer data.

2.1 In the *Main* window, click on + in the toolbar of the *Experiment types* panel and select ***Fingerprint type*** from the list (see Figure 2.1).

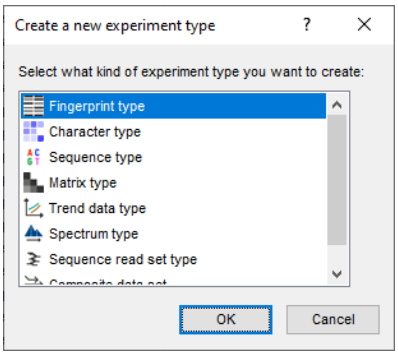

Figure 2.1: The *Create a new experiment type* dialog box.

2.2 Press <***OK***>, enter a name, for example “FragmentAnalyzer” and press <***Next***> (see Figure 2.2).

2.3 In the next window, make sure that ***Two-dimensional TIFF files*** is selected and select the dynamic range of your data (e.g. ***12-bit***). Press <***Next***>.

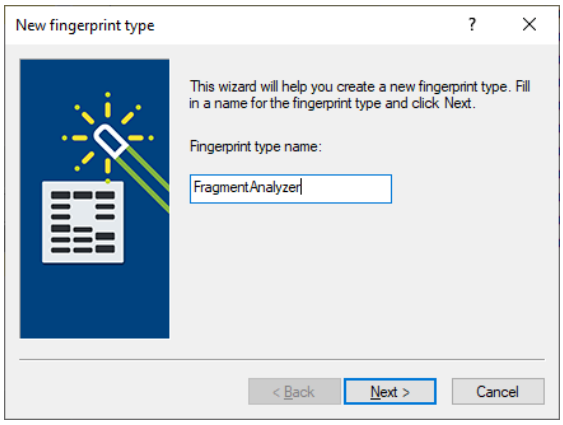

Figure 2.2: Specify a name.

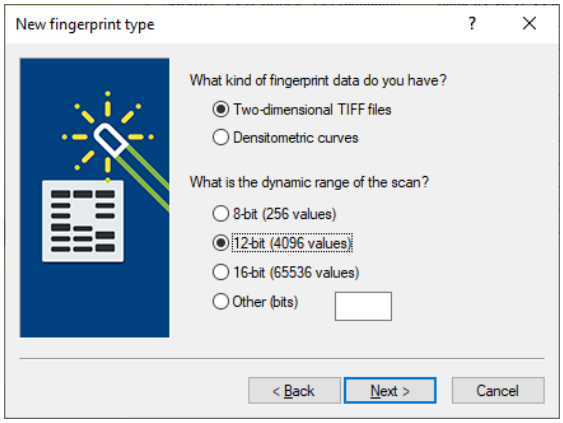

Figure 2.3: Fingerprint settings.

2.4 Press <***Next***> twice without altering the settings and press <***Finish***> to complete the creation of the new fingerprint type.

The *Experiment types* panel now lists the fingerprint type **FragmentAnalyzer**.

2.3 Import routine

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

3.2 With <***Manual selection***> highlighted, press <***Next***> to show all import options in a tree view (see Figure 2.4).

3.3 Select ***Import FragmentAnalyzer curves*** under ***Fingerprint type data*** and press <***Finish***> button to start with the import of the data.

The *Import FragmentAnalyzer data* wizard page appears (see Figure 2.5).

The FragmentAnalyzer import routine accepts FragmentAnalyzer *csv files* and *txt files*.

Pressing the <***Browse***> button allows you to select the csv or text file(s) that you want to import, located on your computer, external drive or on a network location. Alternatively, files can be added to the import list through drag and drop. The number of files and total size is displayed below the list. With the <***Delete***> button all selected files are removed from the import list. All files are

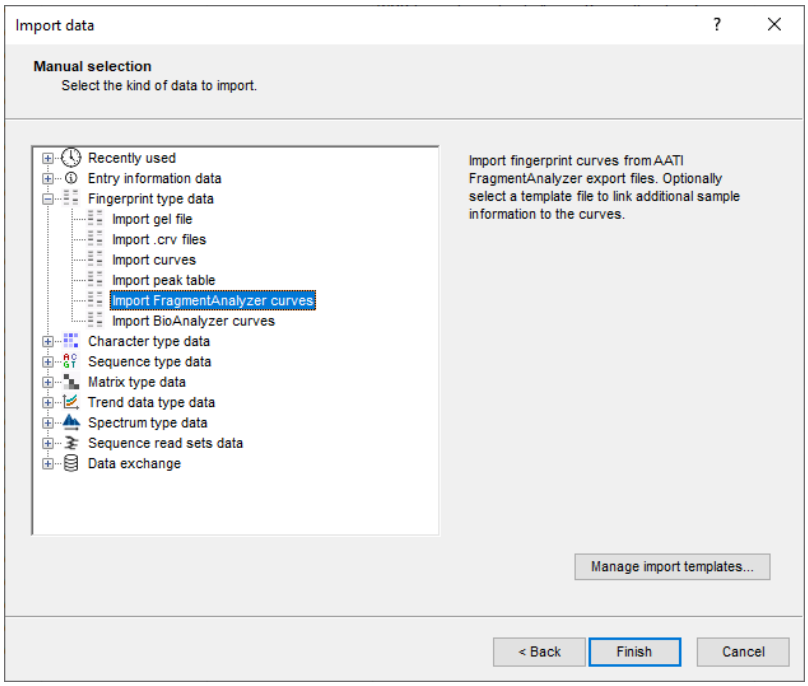

Figure 2.4: The second page of the *Import data* wizard.

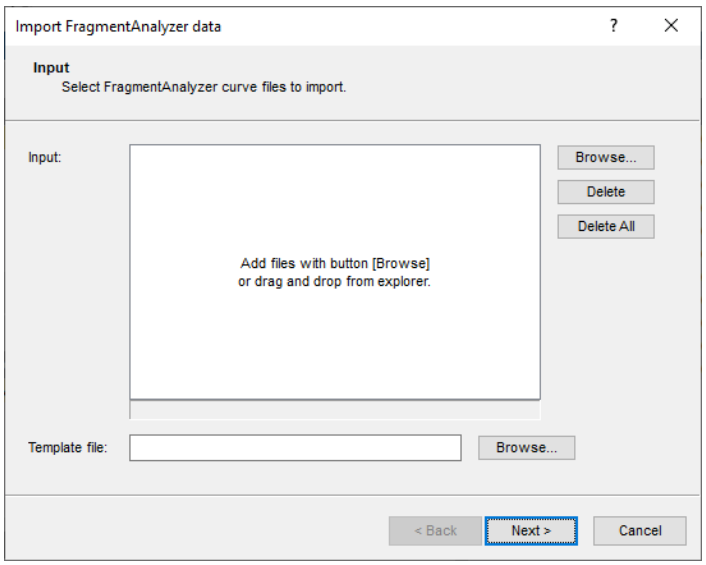

Figure 2.5: The *Import FragmentAnalyzer data* wizard page.

deleted at once from the import list when pressing <***Delete All***>.

Using a *template file* sample information can be imported together with the curves. A *template file* is a simple tab- or comma delimited file with two columns (no headers). The first column should contain the *column names* present in the curve files and the second column should contain the sample information (see Figure 2.6 for an example). In the last step of the wizard, the corresponding BIONUMERICS information field can be selected.

3.4 Browse for the Fragment Analyzer file(s), optionally select a template file and press <***Next***>.

The *Import rules* dialog box is displayed (see Figure 2.7).

The content present in the first row in the csv/txt file corresponds to the first row in the grid. The

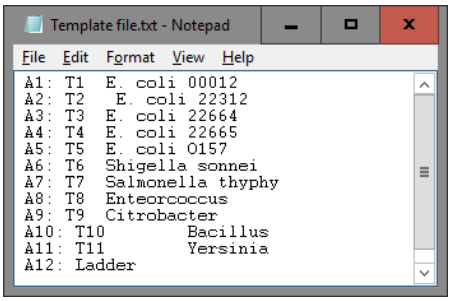

Figure 2.6: A template file: column names (left) and sample information (right).

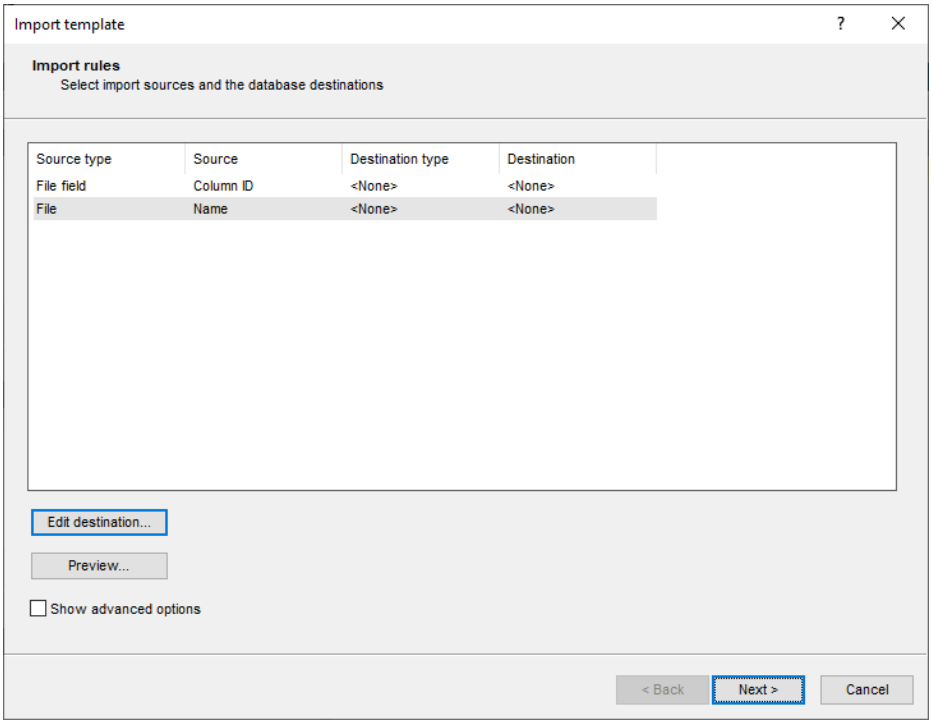

Figure 2.7: The *Import rules* dialog box.

text ***Column ID*** is specified in the ***Source*** column.

Double-clicking on the row opens a new dialog, where the data destination can be selected (see Figure 2.8). Typically, the information in the header row is linked to a new or existing ***Lane info field*** or ***Entry info field***.

Using the last row in the grid, the (parsed) file name of the selected file(s) can be stored in the database. The text ***File*** is specified in the ***Source type*** column and the text ***Name*** is displayed in the ***Source*** column.

3.5 Specify a *destination* for one or more selected rows by pressing the <***Edit destination***> button or by double-clicking (see Figure 2.9 for an example).

3.6 Press <***Preview***> to check the defined rules (see Figure 2.10 for an example). Close the preview.

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

In the *Import curves* wizard page, following settings are available:

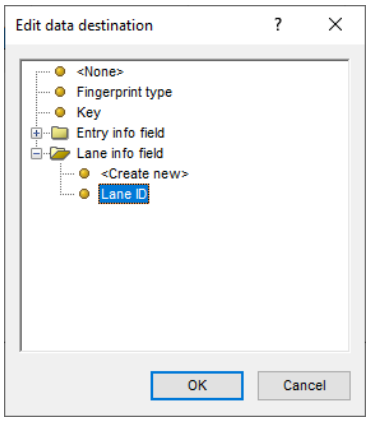

Figure 2.8: Edit data destination.

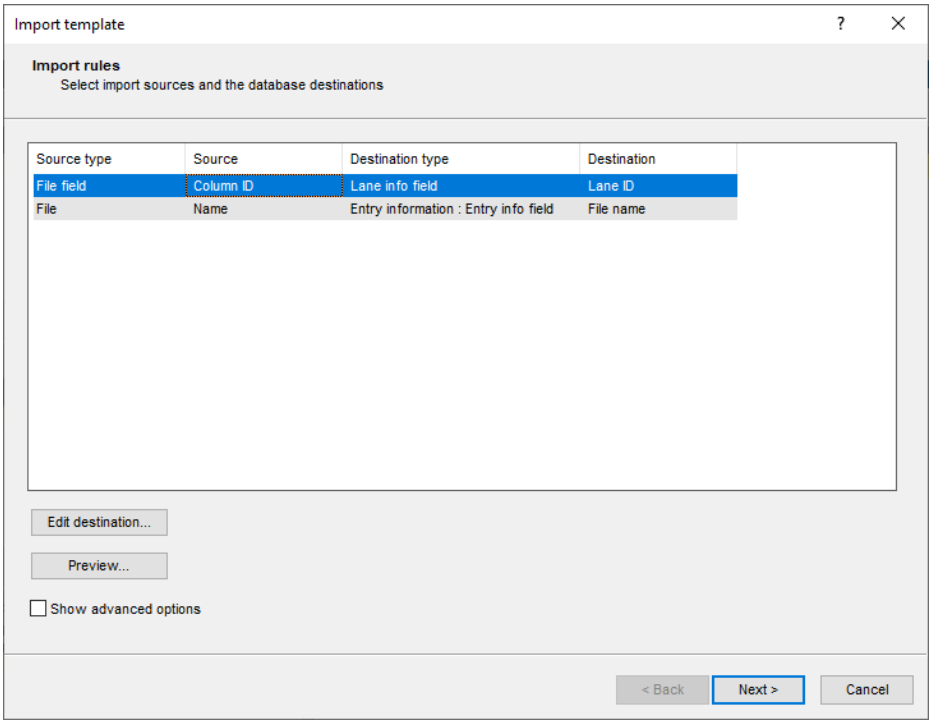

Figure 2.9: Import template.

- The ***Curve start*** and ***Curve stop*** positions are by default set to 0% and 100% respectively. Based on these settings, the complete curves are imported. By entering other start and stop positions as percentages, specific parts of the fingerprint curves can be imported.

- Any peak with a height exceeding the OD range of the fingerprint experiments will appear truncated. To avoid this, a ***Height factor*** can be applied. The ***Height factor*** is by default set to “1.0” (= no height reduction). If the ***Height factor*** is set to e.g. “2”, the heights are reduced by a factor two.

- Checking the option ***Invert curve X-axis*** will invert the curves (fragments at the top of the curve will appear at the bottom and vice-versa) right before they are added to the database.

- With the option ***Do not link standard (ladder) lane*** checked, the ladder lane will not be linked to an entry in the database.

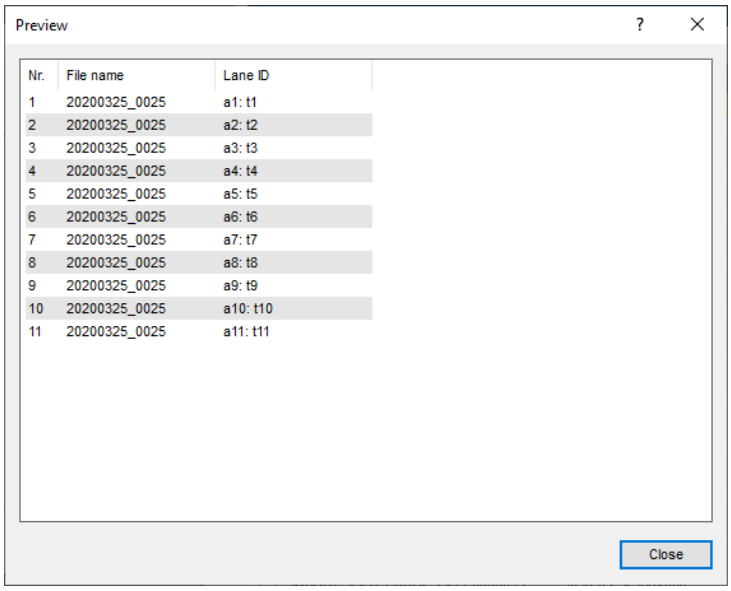

Figure 2.10: Preview.

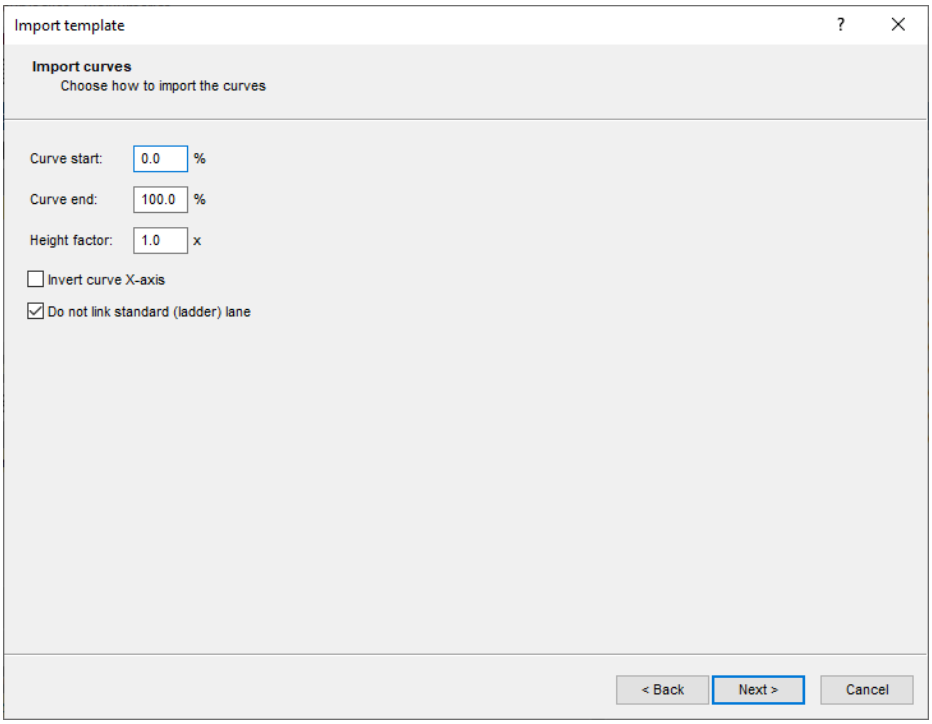

Figure 2.11: The *Import curves* wizard page.

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

The last step prompts for some final settings.

- If a row in the grid is linked to the ***Key*** field in the database, ***Key*** is automatically selected as the entry link field. If entries are already present in the database with the same (parsed) key information, the import tool will link the data to these entries.
- If no row entry in the grid is linked to the ***Key*** field, but one or more rows are linked to an entry information field in the database, these fields can be selected from the list. If entries

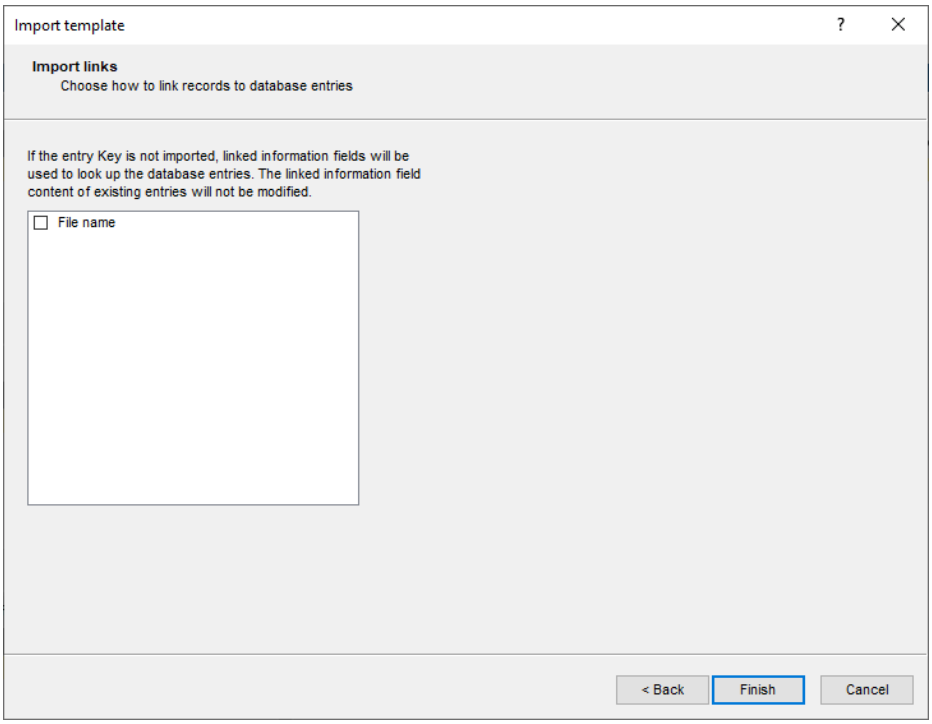

Figure 2.12: The *Import links* dialog box.

are already present in the database with this linked information, the import tool will link the data to these entries. If the entries are not yet present in the database, the data will be linked to new entries in the database (if the option ***Create x entries*** is checked in the last step of the wizard).

- If no fields are selected from the list, no check for existing entries will be performed, and all data will be linked to new entries in the database (if the option ***Create x entries*** is checked in the last step of the wizard). New keys are automatically generated during import.

3.9 Press <***Finish***> to go to the final step.

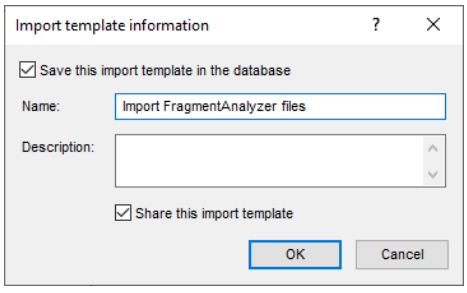

Figure 2.13: The *Import template information* dialog box.

Each import template has its own unique ***Name***. Optionally, a descriptive text string can be entered in the ***Description*** input field.

3.10 Specify a template name (e.g. **Import FragmentAnalyzer files**) and press <***OK***> to save all template settings to the database.

When a template has been created and saved, the template ***Name*** is shown in the *Import templates panel* and is automatically selected (see Figure 2.14). The template ***Description*** is shown

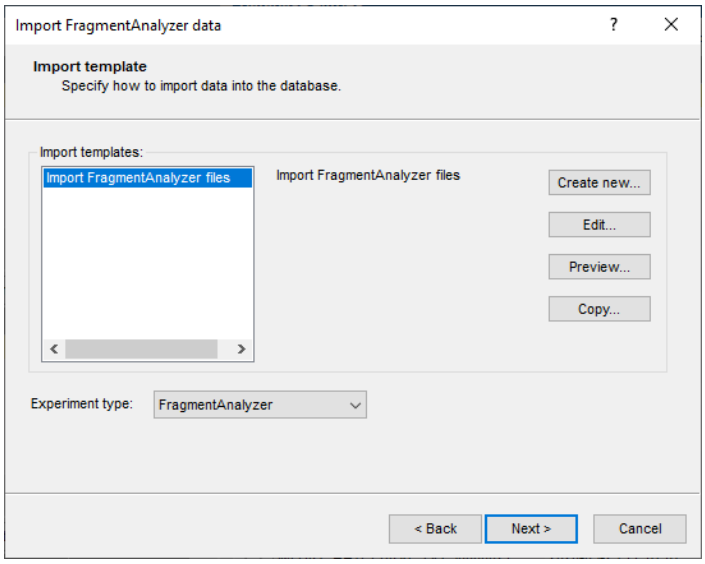

Figure 2.14: Import template

in panel on the right.

The patterns can be linked to an existing fingerprint type experiment or to a new fingerprint type experiment (<***Create New***>). When the fingerprints are linked to a new fingerprint type experiment, the next dialog will prompt for the fingerprint type name (see Figure 2.15). The creation of the new experiment needs to be confirmed (twice).

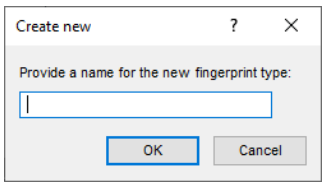

Figure 2.15: Create a new fingerprint type experiment.

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

The last step prompts for some final settings (see Figure 2.16).

- When ***Create x entries*** is checked, the import tool is allowed to create the new entries in the database.
- Check the option ***Update x entries*** if you want the software to be able to update the information for existing entries.
- If the option ***Select modified entries*** is checked, entries in the database that were modified during the import routine will be selected after import.

3.12 Press <***Finish***> to start the import.

Entries are created/updated and are displayed in the *Database entries* panel of the *Main* window. Linked sample information - if defined - is stored in the corresponding entry fields. When the option ***Select modified entries*** was checked, the new/updated entries are marked by a checked ballot box (☑).

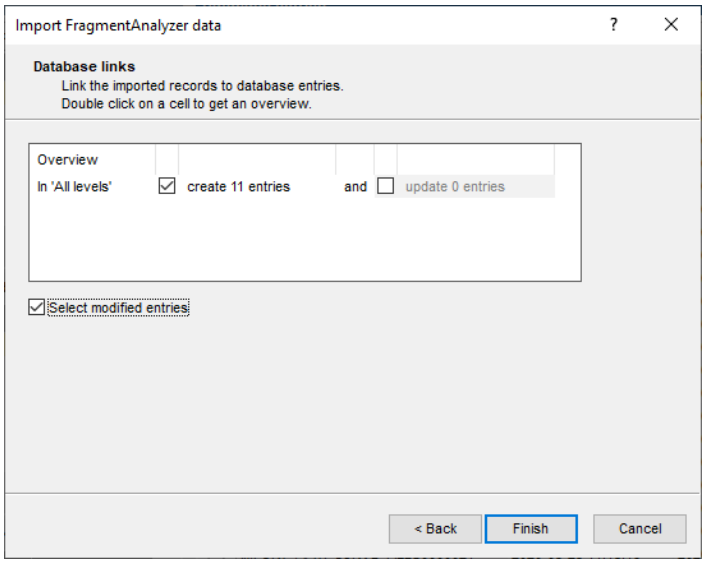

Figure 2.16: Database links.

The fingerprint files are displayed in the *Fingerprint files* panel and contain the corresponding fingerprint patterns. The patterns are linked to the appointed fingerprint type experiment in the database. The presence of a pattern for an entry/experiment combination is indicated with a green colored dot in the *Experiment presence* panel.

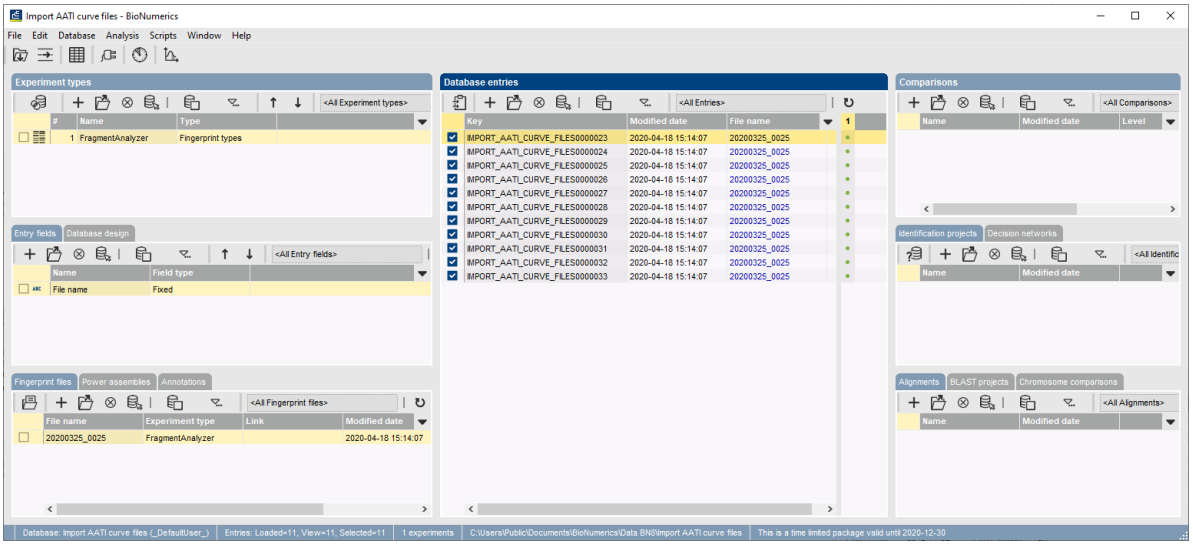

Figure 2.17: The *Main* window after import of FragmentAnalyzer data.

2.4 Processing of FragmentAnalyzer data

4.1 In the *Fingerprint files* panel, double-click on one of the imported files to open the *Fingerprint* window.

4.2 In the *Fingerprint* window, select ***File*** > ***Edit fingerprint data...*** ( ) to open the *Fingerprint processing* window.

The OD range defined for the linked fingerprint type experiment is applied on every imported file.

4.3 To check (and optionally update) the OD range, select ***Edit*** > ***Edit settings...*** ( ) and select the ***Raw data*** tab.

The OD (or intensity) range can be update to any other value if needed: e.g. ***12-bit (4096 values)***, ***16-bit (65536 values)***, ...

4.4 Close the *Fingerprint processing settings* dialog box.

4.5 Make sure the ***Normalization*** tab is selected in the *Fingerprint processing* window.

4.6 Press to enter the normalized view.

For now, the ”normalized view” looks the same as the original view.

4.7 Select the reference lane and press to assign it as a reference lane or select ***References*** > ***Use as reference lane***.

4.8 Right-click on a band in the reference lane and choose ***Add external reference position***.

4.9 Enter the size (in bp) of the reference band. Make sure to only add numerical values (so without bp or kb).

4.10 Repeat this action for all other reference bands in the reference lane.

4.11 Press .

4.12 The program may prompt with the following question: ”The resolution of this gel differs considerably from the normalized track resolution. Do you wish to update the normalized track resolution?” If the question appears, answer <***Yes***>.

The reference system is now defined and saved with the fingerprint type experiment. When processing new fingerprint files, run with the same reference system, the assignment of the reference positions can be skipped: just assign the correct lane as reference lane (***References*** > ***Use as reference lane***), execute ***Normalization*** > ***Auto assign...*** ( ) and inspect the assignments made.

4.13 Select the ***Bands*** tab in the *Fingerprint processing* window.

If you want to use the curves to compare the patterns, no bands need to be assigned in the sample lanes. If you want to compare the patterns using bands, you will need to assign bands in the sample lanes in the last step. Usually, assigning bands in the sample lanes is done first with the software's automatic band search, followed by manual corrections. Some trial and error might be required to find the best settings for the automatic band search.

4.14 To automatically search for bands, press or select ***Bands*** > ***Auto search bands***.

In the *Band search* dialog box, the currently selected lane is shown along the bottom.

4.15 To scroll through other lanes, press the < and > buttons on the left and right sides of the curve.

4.16 Press <***Search on all lanes***> to execute the band search with the specified settings.

Bands that are found are marked with a green horizontal line. Incorrect band assignments can be edited manually. For more information we refer to the reference manual.

4.17 After you are satisfied with the band assignments, press to save the file.

4.18 Exit the *Fingerprint processing* window by selecting ***File*** > ***Exit***.

When a reference system is added to a fingerprint type experiment, a calibration curve needs to be created for this reference system that translates all band positions into metrics.

4.19 Open the fingerprint experiment type (**FragmentAnalyzer** in this workflow) by double-clicking on the experiment type in the *Experiment types* panel.

4.20 In the *Fingerprint type* window, select ***Settings*** > ***Edit reference system*** or double-click on **R01**.

The *Fingerprint Reference system* window appears with the message: "Could not calculate calibration curve. Not enough markers."

4.21 When the molecular weights were entered as names for the reference positions, the molecular weights can be copied by selecting ***Metrics*** > ***Copy markers from reference system...***. Confirm the action.

4.22 Designate a metric unit with ***Metrics*** > ***Assign units...*** enter e.g. **kb** and press <***OK***>.

4.23 Close the *Fingerprint Reference system* window, and close the *Fingerprint type* window.

More information about the follow-up analysis tools that can be applied on the patterns can be found in the tutorials on the BIONUMERICS website or in the reference manual.

Chapter 3

Import of Agilent 2100 Bioanalyzer curve files

3.1 Data format

The 2100 Bioanalyzer is a capillary electrophoresis instrument from the company Agilent (https://www.agilent.com/). The generated electropherograms can be exported from the Agilent 2100 Bioanalyzer software in csv format and imported in BIONUMERICS after installation of the *Import fingerprint tools plugin*.

In the "Electrophoresis Export Options" dialog of the Agilent 2100 Bioanalyzer software (see Figure 3.1), make sure to check **Sample data** ("Creates csv files containing the sample data, one file per sample"). It does not matter whether **aligned sample data** or **unaligned sample data** are exported, since the data are normalized in BIONUMERICS anyway. **Result tables** need not to be checked.

Each exported csv file contains the full densitometric curve for a single profile, with some additional meta data.

3.2 Setting up an Agilent 2100 Bioanalyzer fingerprint type experiment

Initially, a new fingerprint type experiment needs to be created which will hold the imported Agilent 2100 Bioanalyzer curve data.

2.1 In the *Main* window, click on + in the toolbar of the *Experiment types* panel and select ***Fingerprint type*** from the list (see Figure 3.2).

2.2 Press <***OK***>, enter a name, for example "Agilent 2100 Bioanalyzer" and press <***Next***> (see Figure 3.3).

2.3 In the page of the *New fingerprint type* dialog box, make sure that ***Two-dimensional TIFF files*** is selected and select the dynamic range of your data (e.g. ***8-bit***). Press <***Next***>.

2.4 Press <***Next***> twice without altering the settings and press <***Finish***> to complete the creation of the new fingerprint type.

The *Experiment types* panel now lists the fingerprint type **Agilent 2100 Bioanalyzer**.

The import routine assumes that two reference positions are defined, corresponding to the upper

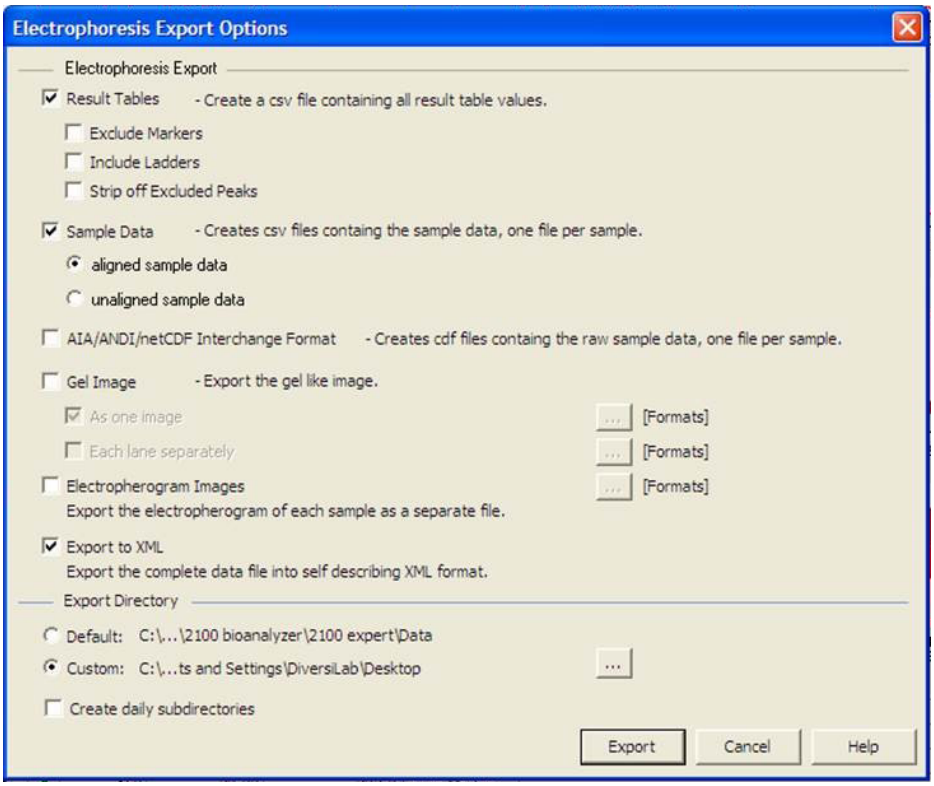

Figure 3.1: Electrophoresis Export Options dialog of the Agilent 2100 Bioanalyzer software.

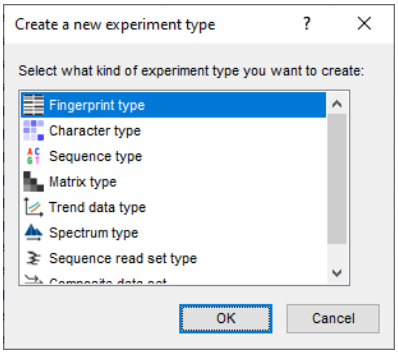

Figure 3.2: The *Create a new experiment type* dialog box.

and the lower marker. These reference positions can be created manually in the *Fingerprint type* window via ***Settings*** > ***New reference system (positions)...*** or based on actual band positions in the *Fingerprint processing* window during the import of the first fingerprint file.

3.3 First-time import: creating an import template

The first time that Agilent 2100 Bioanalyzer csv files are imported in a BIONUMERICS database, an *import template* should be created. The data files contain fingerprint curves as well as meta data and the import template determines where (which fields, entries, experiments, etc.) this information will be stored in the BIONUMERICS database. After the initial import, each subsequent

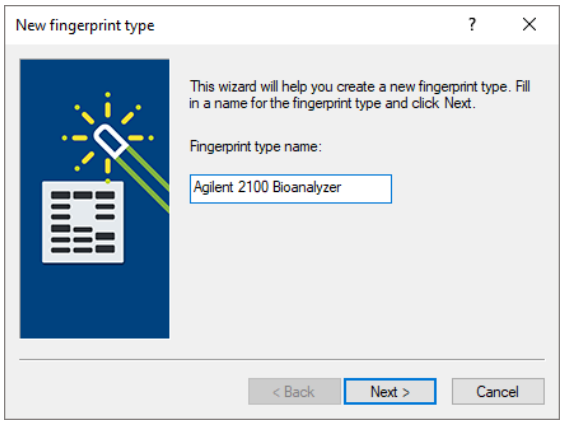

Figure 3.3: Specify a name.

import action (see 3.5) will be much faster, since a saved import template can simply be re-used.

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

3.2 With <***Manual selection***> highlighted, press <***Next***> to show all import options in a tree view (see Figure 3.4).

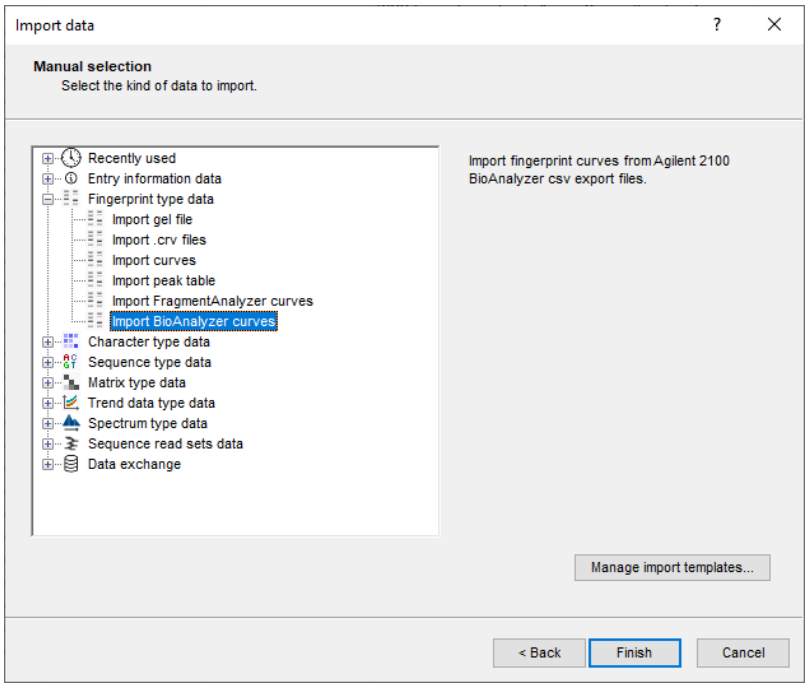

Figure 3.4: The second page of the *Import data* wizard.

3.3 Select ***Import Bioanalyzer curves*** under ***Fingerprint type data*** and press <***Import***> to start with the import of the data.

The *Import Agilent 2100 Bioanalyzer data* wizard page appears (see Figure 3.5).

The Agilent Bioanalyzer import routine only accepts *csv files* (see 3.1).

Pressing the <***Browse***> button allows you to select the csv files that you want to import, located on your computer, external drive or on a network location. Alternatively, files can be added to the import list through drag and drop. The number of files and total size is displayed below the list.

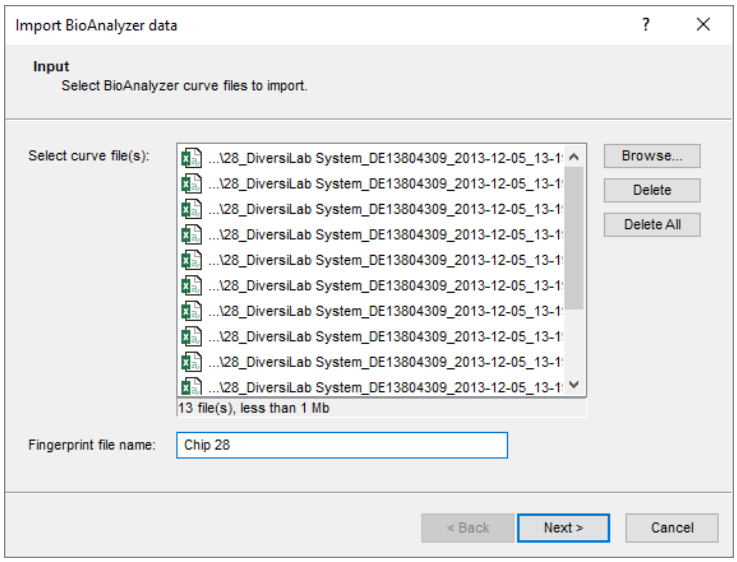

Figure 3.5: The *Import Agilent 2100 Bioanalyzer data* wizard page.

With the <***Delete***> button, selected files are removed from the import list. All files are deleted at once from the import list when pressing <***Delete All***>.

Since a separate fingerprint file will be created for each import batch, it makes sense to select all samples from a single chip.

The suggested ***Fingerprint file name*** corresponds to the folder name, but any other name can be entered in the text box.

Do not select any `*_Results.csv` files, as these have a very different format from the curve data files (`*_Ladder.csv` and `*_Sample.csv` files).

3.4 Browse for the Agilent 2100 Bioanalyzer files, enter a different ***Fingerprint file name*** if needed and press <***Next***>.

The *Import rules* dialog box is displayed (see Figure 3.6).

The *Import rules* dialog box lists the information present in the selected files as **Source**, their linked **Source type** and the **Destination** component they are associated with (initially all set to <None>).

Agilent 2100 Bioanalyzer csv files provide meta data as they were entered in the Agilent software. Each piece of information is presented as a ”File field”, which can be linked to a destination: either an entry information field or a fingerprint information field in BIONUMERICS.

For example, if we want to store all meta data from the csv files into entry information fields in BIONUMERICS:

3.5 Click on the first file field in the list (”Data File Name”) to highlight it. While holding the **Shift**-key, click on the last file field (”Number of Events”) to highlight all file fields.

3.6 Press the <***Edit destination...***> button.

3.7 In the *Edit data destination* dialog box, click on ***Entry info field*** and press <***OK***>.

If the corresponding information fields are not yet present in your database, they need to be created first.

3.8 Press <***OK***> in the *Create new* dialog box and confirm the creation of the entry info fields.

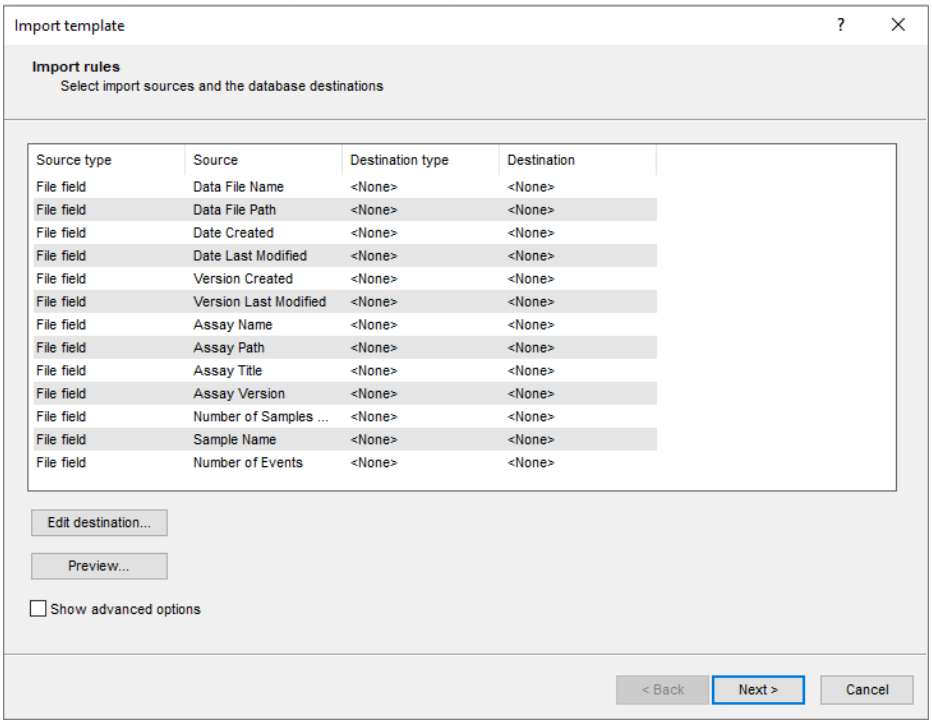

Figure 3.6: The *Import rules* dialog box, displaying import rules for Agilent 2100 Bioanalyzer csv files.

In case the file name contains useful information, this data source can be used after checking ***Show advanced options*** and pressing <***Add rule...***> in the *Import rules* . dialog box

3.9 Press <***Next***> in the *Import rules* dialog box.

This action displays the *Import curves* wizard page (see Figure 3.7).

Following settings are available from this dialog box:

- The ***Curve start*** and ***Curve stop*** positions are by default set to 0% and 100% respectively. Based on these settings, the complete curves are imported. By entering other start and stop positions as percentages, specific parts of the fingerprint curves can be imported.
- Any peak with a height exceeding the OD range of the fingerprint experiments will appear truncated. To avoid this, a ***Height factor*** can be applied. The ***Height factor*** is by default set to “1.0” (= no height reduction). If the ***Height factor*** is set to e.g. “2”, the heights are reduced by a factor two.
- Checking the option ***Invert curve X-axis*** will invert the curves (fragments at the top of the curve will appear at the bottom and vice-versa) right before they are added to the database.

In most import scenarios, these settings can be left to their default values.

3.10 Press <***Next***> in the *Import curves* wizard page to accept the default settings.

The *Import links* dialog box appears, listing the fields from the Agilent 2100 Bioanalyzer csv files that are imported into entry information fields. Since none of the fields contain a unique identifier for the imported samples, no fields should be linked to let BIONUMERICS automatically create unique entry keys.

3.11 Press <***Finish***>.

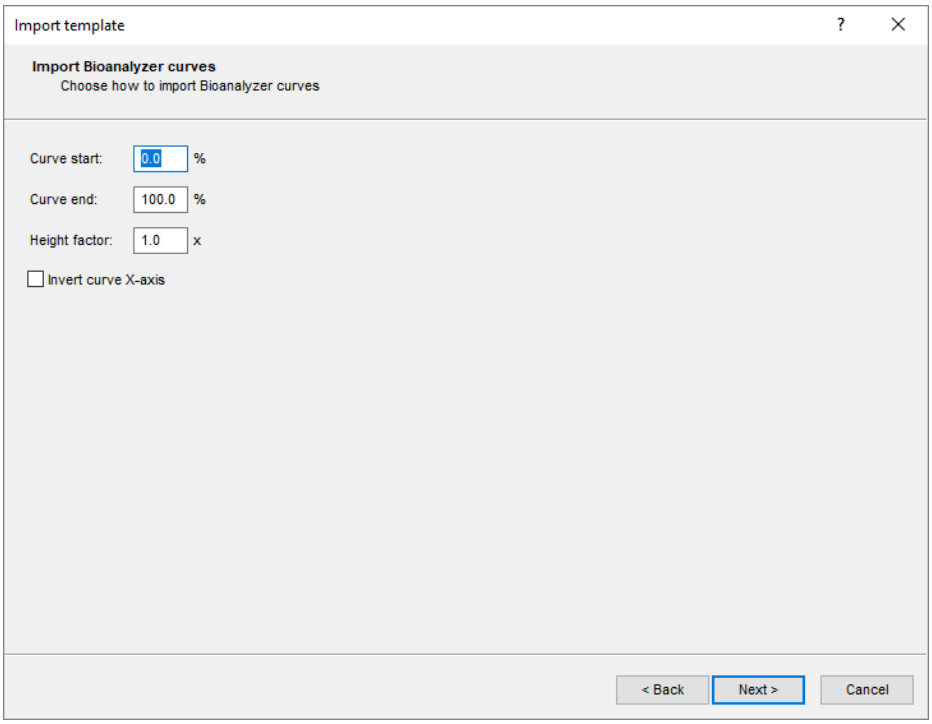

Figure 3.7: The *Import curves* wizard page.

3.12 Enter a ***Name*** (e.g. “Agilent CSV”) for the newly created import template and optionally a ***Description***.

3.13 Press <***OK***> to save the import template in the database.

The *Import template Bioanalyzer data* wizard page appears (see Figure 3.8).

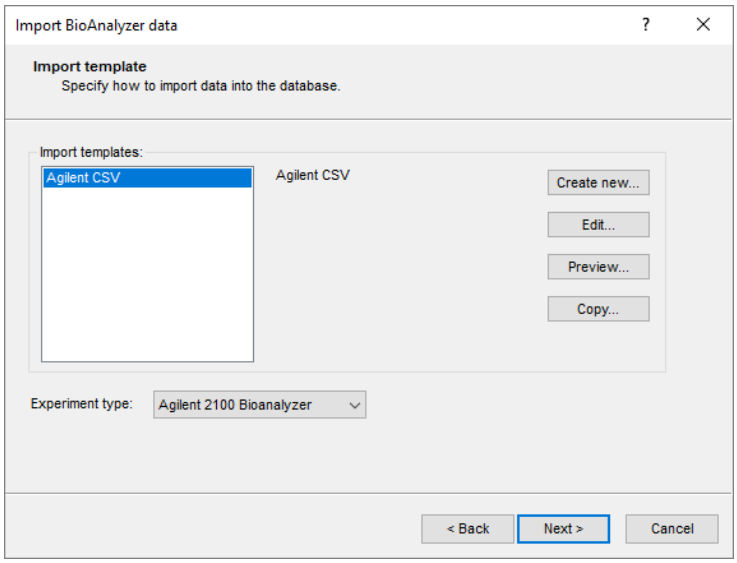

Figure 3.8: The *Import template Bioanalyzer data* wizard page.

When an import template has been created and saved, the template ***Name*** is shown in the *Import templates panel* and is automatically selected (see Figure 3.8). The template ***Description*** is

shown in panel on the right.

3.14 With the **Agilent CSV** import template highlighted and the corresponding ***Experiment type*** selected (see 3.2), press <***Next***>.

The last page in the wizard indicates how many entries will be created and/or updated in this import action. The import template that was created in this example will always add new entries and never update existing ones, since a unique entry key is generated for each profile.

3.15 Press <***Finish***> to import the csv files into the database.

Each csv file corresponds to a curve and is imported as a lane of a fingerprint file, under the name specified during import. Each lane is linked to a database entry, for which several information fields are entered. All curves from an import batch are grouped into a single fingerprint file. The fingerprint file is automatically normalized based on the upper and lower markers in each profile and a band search is performed.

After the first import, the warning message ”No reference system found. Please assign reference positions manually.” will pop up when the fingerprint type does not have a reference system. See 3.4 on how to create a reference system based on the upper and lower marker in an actual fingerprint profile.

3.4 Creating a reference system

To be able to perform its automatic normalization, a reference system needs to be present in the fingerprint type with two reference positions corresponding to the upper and lower marker, respectively. Such a reference system can be created based on a representative fingerprint profile. This needs only be done for the first chip imported; the reference system will be used for any subsequent chip.

4.1 In the *Fingerprint processing* window of the first imported chip, choose a suitable pattern (e.g. a molecular weight ladder). The overall intensity of the profile should be sufficient and the upper and lower marker unambiguously identifiable.

4.2 Show the gel image in normalized view with ***Normalization*** > ***Show normalized view*** (, **Shift+N**).

4.3 Click on the band corresponding to the upper marker in the chosen profile and select ***References*** > ***Add external reference position***.

4.4 In the dialog that appears, enter “7000” as band name and press <***OK***>.

4.5 Scroll down in the gel image and click on the band corresponding to the lower marker in the same profile.

4.6 Select ***References*** > ***Add external reference position***, enter “150” as band name and press <***OK***>.

This creates the necessary reference system. Since automatic normalization was not possible during the import of this first chip, we need to normalize the gel manually.

4.7 With the reference position **150** highlighted, **Ctrl+click** on the lower marker band in each profile to assign the band to this reference position.

4.8 Click on **7000** to make it the active reference position and **Ctrl+click** on the upper marker in each profile.

4.9 Update the normalization with ***Normalization*** > ***Update normalization*** (, **Ctrl+U**).

Next, we will assign bands on the fingerprint profiles. This step is not required when only curve-based similarity coefficients will be used to compare profiles.

4.10 Select ***File*** > ***Next step*** ( ) to proceed to the next step.

4.11 Select ***Bands*** > ***Auto search bands*** ( ) and press <***Search on all lanes***>.

4.12 Check the band assignments.

If too many band assignments are wrong, repeat the automatic band search with different settings. Minor tweaks to the band assignment can be done manually.

4.13 Save your work with ***File*** > ***Save*** (, **Ctrl+S**).

A warning may appear about the normalized gel resolution. It is safe to update the normalized track resolution to the actual resolution of the gel.

4.14 Press <***Yes***> in the confirmation message.

4.15 Close the *Fingerprint processing* window with ***File*** > ***Exit***.

3.5 Routine import

Once an import template is created (see 3.3) and reference system is present in the fingerprint type (see 3.4), the routine import of Agilent 2100 Bioanalyzer files is very quick.

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

5.2 Press the <***Browse***> button and browse for the CSV files that you want to import. Alternatively, add the files from an Explorer window through drag and drop.

5.3 With the ***Import Bioanalyzer curves*** option highlighted, press <***Finish***>.

5.4 Enter a different ***Fingerprint file name*** if needed and press <***Next***>.

5.5 Press <***Next***> in the *Import template Bioanalyzer data* wizard page.

5.6 Press <***Finish***> to start the import.

⚠ In case one or more ”Failed to find start and/or end peak on lane x.” errors appear, the mentioned lanes should be normalized manually to avoid errors during comparison. See 3.6 for instructions.

3.6 Manually correcting normalization issues

Whenever a lane cannot be automatically normalized during import, an error ”Failed to find start and/or end peak on lane x.” is produced. This error should not be ignored: since there are only two reference positions, the normalization will be completely off when either the start or stop peak was not found.

In most cases, the issue can be fixed by manually assigning peaks to reference positions:

6.1 In the *Fingerprint files* panel of the *Main* window, click on the fingerprint file that has the issue.

6.2 Select ***Open fingerprint data...*** ( ) to open the fingerprint file in the *Fingerprint processing* window.

6.3 In the bottom of the window, click on the *Normalization* tab or select ***File*** > ***Next step*** ( ) twice to show the normalization step.

By default, the reference position that corresponds to the upper marker (named **7000**) will be active. If the upper marker was not identified correctly in a lane, a band will instead be (wrongly) assigned at the exact reference position (see Figure 3.9 for an example). This needs to be corrected manually.

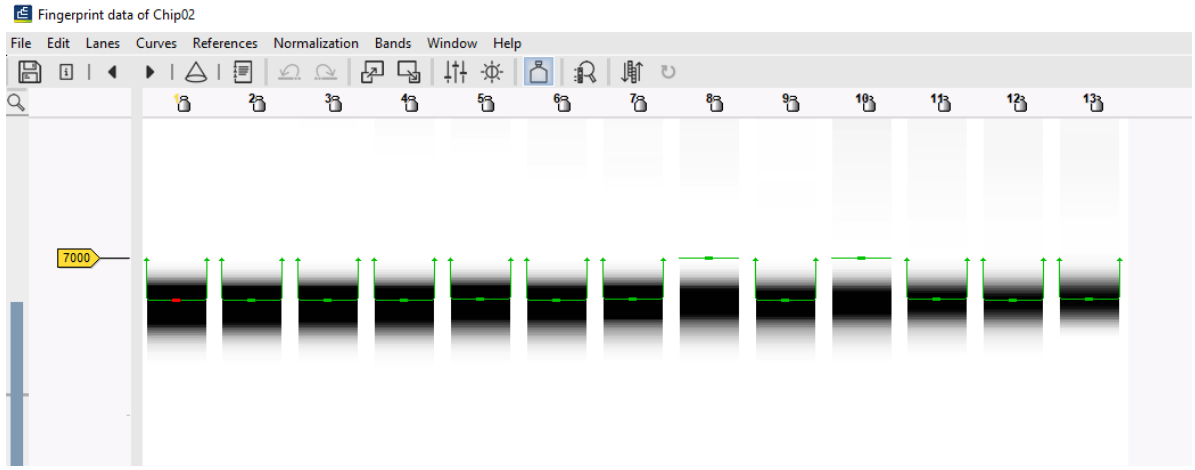

Figure 3.9: Example of a normalization issue: the upper marker was not found on lanes 8 and 10.

6.4 In the first lane with an issue, **Ctrl+Click** on the position of the upper marker to link that peak to the reference position.

6.5 Repeat the previous step for any lane in which the upper marker was not detected.

Next, check for any issues with the lower marker:

6.6 Scroll down the gel image and click on reference position **150**, corresponding to the lower marker.

6.7 **Ctrl+Click** on the position of the lower marker in any lane where the lower marker went undetected.

6.8 Finally, save the edits with ***File*** > ***Save*** (, **Ctrl+S**) and close the *Fingerprint processing* window.

Chapter 4

On-the-fly correction of curve alignments

4.1 Introduction

In order to correct for the normalization problems associated with having only two markers, the plugin can perform an on-the-fly alignment of the densitometric curves present within a comparison. The alignment algorithm takes the similarity matrix as a guide to calculate a weighted average curve for each trace within the set. The average curve is based on all curves within the set, but since it is weighted according to the similarity, curves that are more similar to the trace contribute heavier to the average curve. Each trace's curve is then aligned to its weighted average profile. As a result, traces are particularly aligned to other traces that are highly similar based upon the non-corrected curves. Therefore, in order to obtain the best results with the alignment algorithm, it is recommended to apply a large optimization value, e.g. 3%, while calculating the correlation-based dendrogram. This alignment can be done in two ways:

1. **Non-linear shift with fixed edges** (see Figure 4.1): A non-linear shift is performed on the curve to obtain the highest correlation between the trace and its weighted average. The extremes of the curve are thereby fixed as anchor points so that a global stretch/compression is not possible. The shift is based on a quadratic function with one degree of freedom

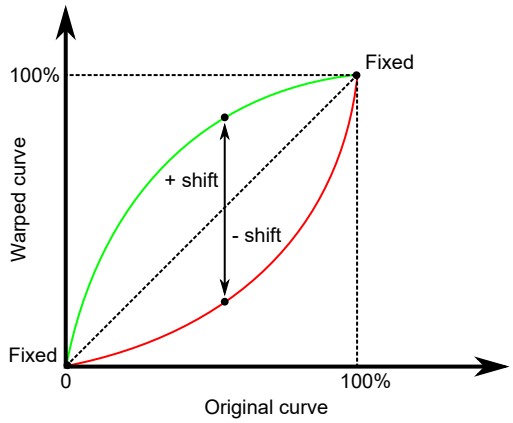

Figure 4.1: Non-linear shift with fixed edges.

and requires no time consuming calculations. Considering the fact that the extremes of the curves correspond to the marker peaks used for the normalization performed by the

DiversiLab software, this alignment can be seen as meaningful. In practice, it will correct most of the distortion in the traces. However, it is frequently observed that the distortion is not maximal in the center of the traces, so that bands towards the edges might still be not well-aligned.

2. **Global shift with linear stretch/compression** (see Figure 4.2): The curve is aligned to its weighted average by means of a global shift and a linear stretch/compression factor. The extremes of the curves are thereby not used as fixed anchor points. This alignment has

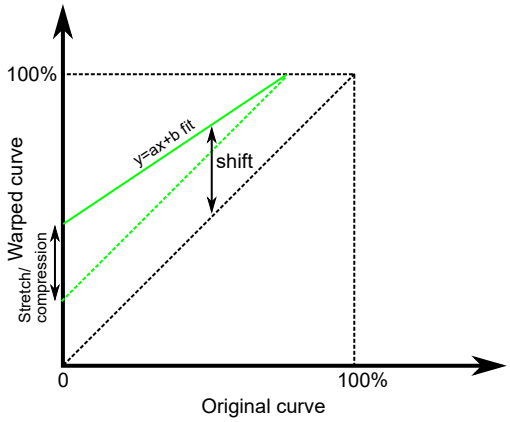

Figure 4.2: Global shift with a linear stretch or compression.

two degrees of freedom and is therefore slower that the non-linear shift with fixed edges. It will usually result in a better alignment of all major bands in the patterns. However, because of the greater freedom, some ”overcorrection” might occur. The alignment is only shown on the densitometric curves, **not** on the gel strips. It is important to realize that the alignment is based on the chosen set of entries within the current comparison. By leaving out or adding entries to the comparison, other alignments might be obtained. The alignment is therefore not saved in the database.

4.2 Calculating the alignment

The auto-correction can only be done if a similarity matrix is present for the fingerprints. The alignment happens on-the-fly on densitometric curves loaded in the comparison. You should therefore display the densitometric curves prior to running the alignment algorithm:

2.1 Select ***Fingerprints*** > ***Show densitometric curves*** ( ) in the *Comparison* window.

2.2 Select ***Fingerprints*** > ***Auto-correct curves*** to perform the auto-correction.

Four settings need to be provided (see Figure 4.3).

- ***Keep edges of fingerprint active zone fixed***: if this setting is enabled, no global stretch or compression of the curve is possible within the fingerprint active zone (see Figure 4.1). If this setting is switched off, a linear stretch/compression is applied on the curve (see Figure 4.2).
- ***Maximum shift*** (as % of the curve length).
- ***Maximum stretch/compression*** (as % of the curve length).

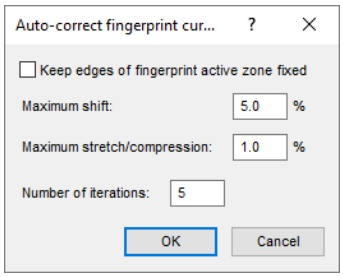

Figure 4.3: The *Auto-correct fingerprint curves* dialog box.

- ***Number of iterations***: this value should be chosen in function of the allowed shift and stretch/compression. More iterations are required to perform larger shifts and stretches/-compressions.

If a linear stretch/compression is applied (no fixed edges), the stretch/compression value can still be set to 0%. In that case, only a shift between the traces is possible, similar as what is achieved using the ***Optimization*** parameter in the clustering wizard (see Figure 4.4).

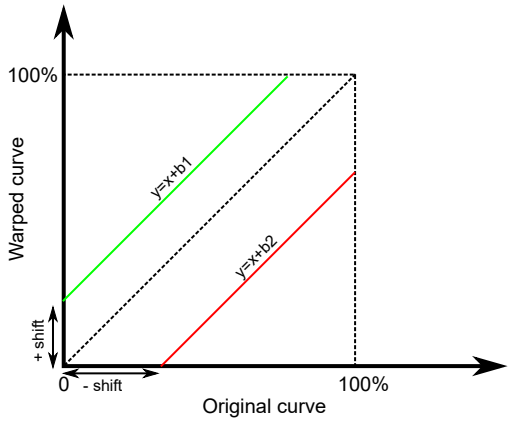

Figure 4.4: Linear stretch or compression applied, with a shift only.

2.3 As a first example, check ***Keep edges of fingerprints fixed***. Use 5.0% as ***Maximum shift*** and enter “5” as ***Number of iterations***. Press <***OK***>. The resulting image looks as in Figure 4.5:

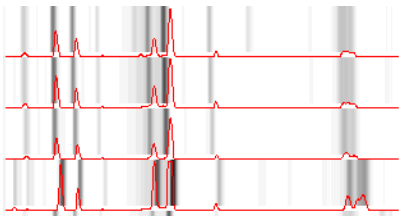

Figure 4.5: Detail

The gel strips display the original traces, whereas the curves are corrected by the auto-correct tool. This viewing mode can be useful to monitor the settings and the results of the alignment algorithm. To display the corrected curves as gel strips rather than densitometric curves, proceed as follows:

2.4 In the *Main* window, double-click on the **Agilent 2100 Bioanalyzer** fingerprint type to open it.

2.5 In the *Fingerprint type* window, select ***Layout*** > ***Show curves as images*** and close the window.

2.6 In the *Comparison* window, click inside the *Experiment data* panel to update the image and select ***Layout*** > ***Show image*** ( ) to hide the original gel images. The corrected curves are now shown as gel strips.

2.7 You can repeat steps Instruction 2.2 to Instruction 2.3 with ***Keep edges of fingerprints fixed*** switched off and 1.0% as ***Maximum stretch/compression*** to see the result of the alignment including stretch/compression.

Figure 4.6 illustrates the effect of the alignment without stretch/compression and with stretch/compression, respectively.

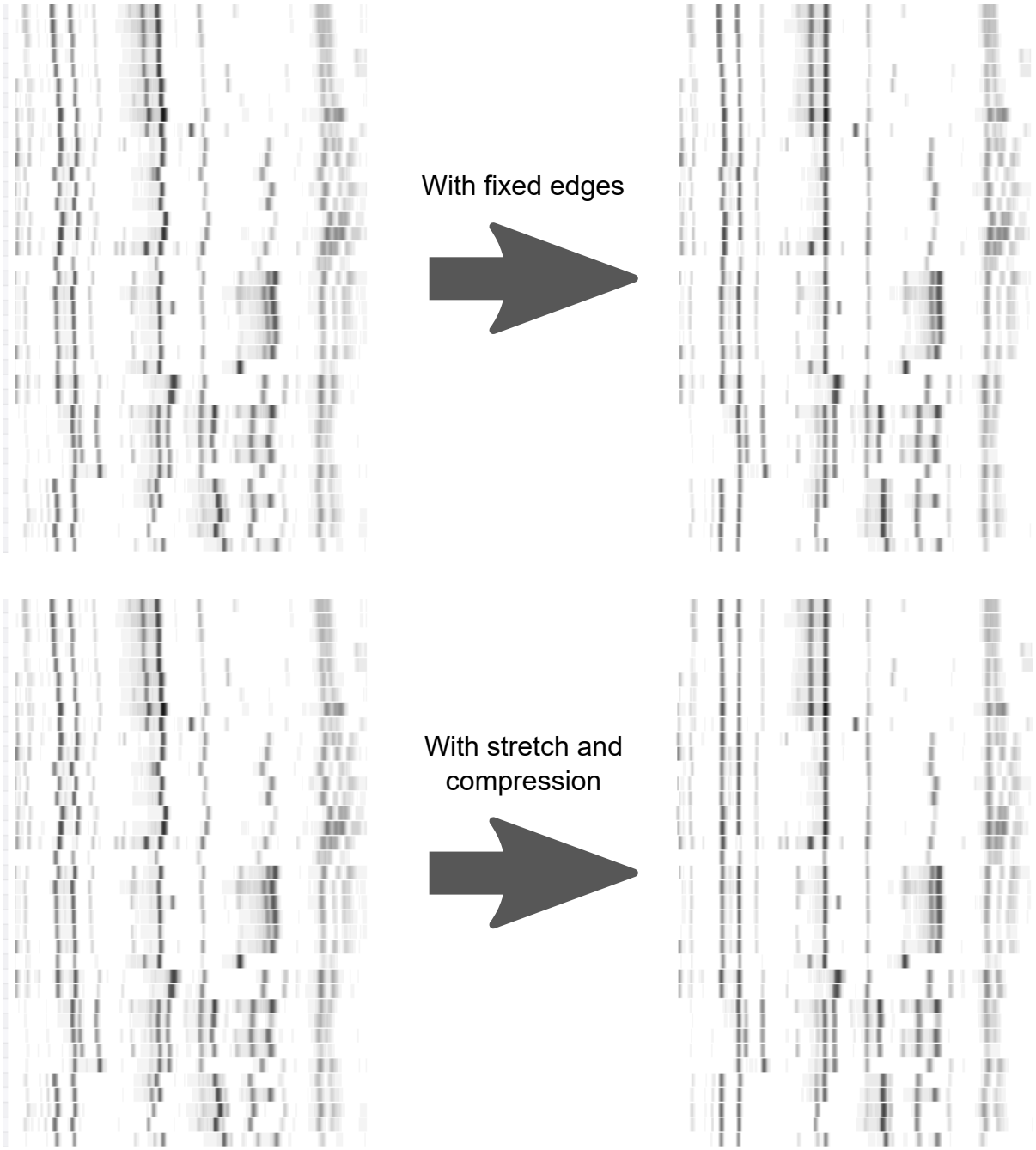

Figure 4.6: Effects of the different alignment options.

The auto-correct function should be used carefully and the result should always be compared with the original traces to verify that the algorithm did not perform excessive shift, stretching or compression on (groups of) patterns. It is recommended to use the function at first with very low values for shift and stretch/compression, and gradually increase these until the results are satisfactory without observing ”over-corrected” traces.

2.8 A clustering can now be calculated on the aligned curves. If the auto-correct algorithm performed additional correction to the traces that could not be obtained by the ***Optimization*** function, the clusters will become more homogeneous.

To show the improved normalization in the gel images, select ***Layout*** $>$ ***Show curves as images*** in the *Fingerprint type* window prior to applying the auto-correction.

Since the alignment is calculated on densitometric curves loaded in the comparison, hiding the curves and showing them again using ***Fingerprints*** $>$ ***Show densitometric curves*** () causes the alignment to disappear.

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