BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

DiversiLab plugin

PLUGINS

VERSION 7.6

www.applied-maths.com

Contents

NOTES

SUPPORT BY APPLIED MATHS

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

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

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

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

LIMITATIONS ON USE

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

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

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

Chapter 1

Starting and setting up BioNumerics

1.1 Introduction

This plugin is used to import and analyze densitometric traces from the DiversiLab system (bioMérieux SA). The traces are exported as XML reports from the web-based DiversiLab software, containing normalized curves and strain information.

1.2 Startup program

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

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

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

1.3 Creating a new database

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

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

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

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

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

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

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

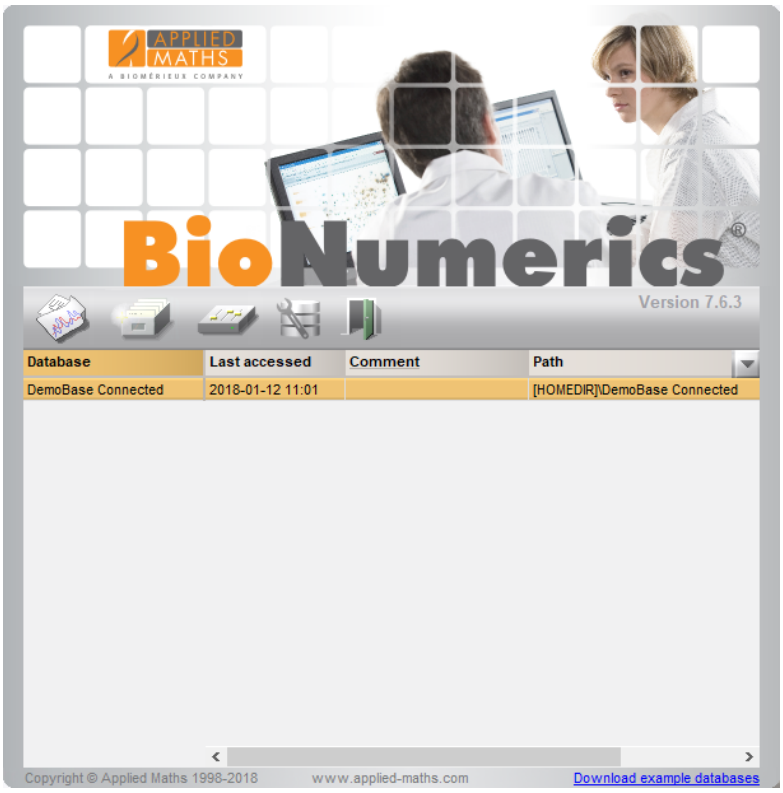

Figure 1.1: The *BioNumerics Startup* window.

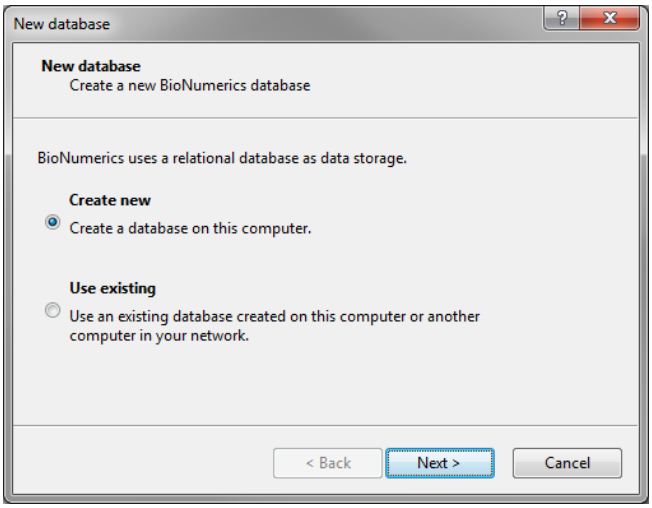

Figure 1.2: The *New database* wizard page.

The *Plugins* dialog box appears.

1.4 Installing the DiversiLab plugin

If a database is opened for the first time, the *Plugins* dialog box will appear by default (see Figure 1.4).

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

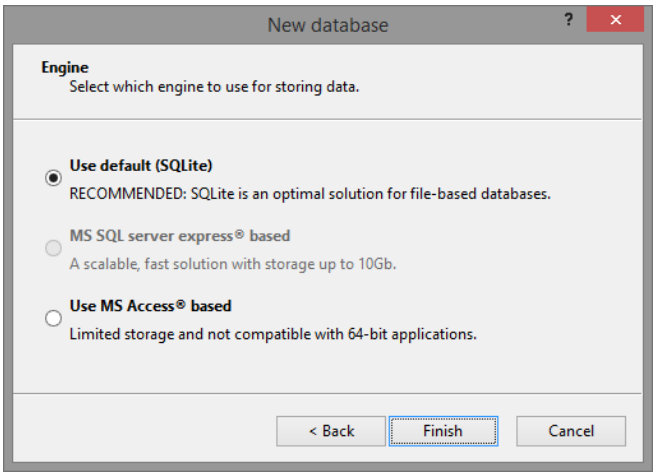

Figure 1.3: The *Database engine* wizard page.

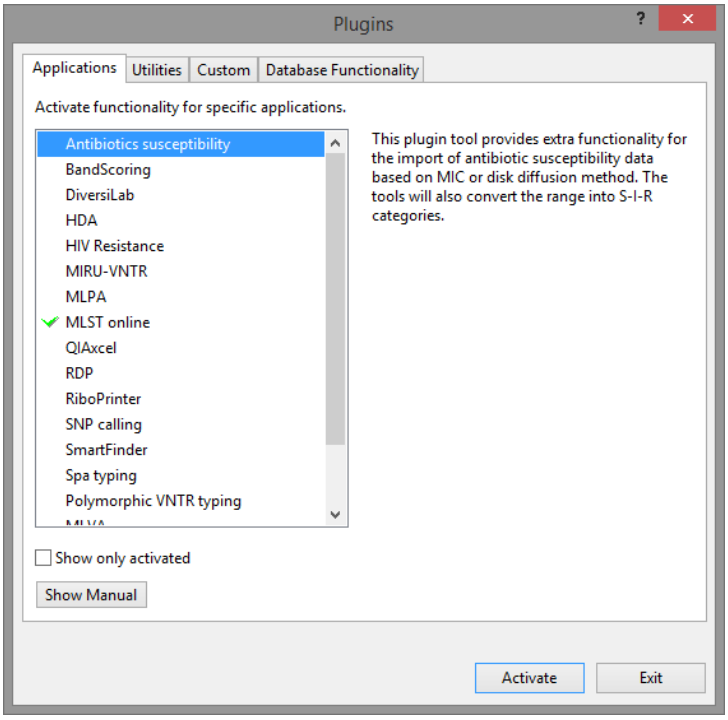

Figure 1.4: The *Plugins* dialog box.

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

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

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

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

4.1 Select the *DiversiLab plugin* from the list in the *Applications tab* and press the <***Activate***> button.

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

4.3 Press <***Proceed***> (or <***Exit***>) to close the *Plugins* dialog box and to continue to the *Main* window.

4.4 Close and reopen the database to activate the features of the *DiversiLab plugin*.

The ***Import DiversiLab XML files*** item is activated in the *Import* dialog box (see Figure 1.5). This dialog is called when selecting ***File*** > ***Import...*** (, **Ctrl+I**) in the *Main* window.

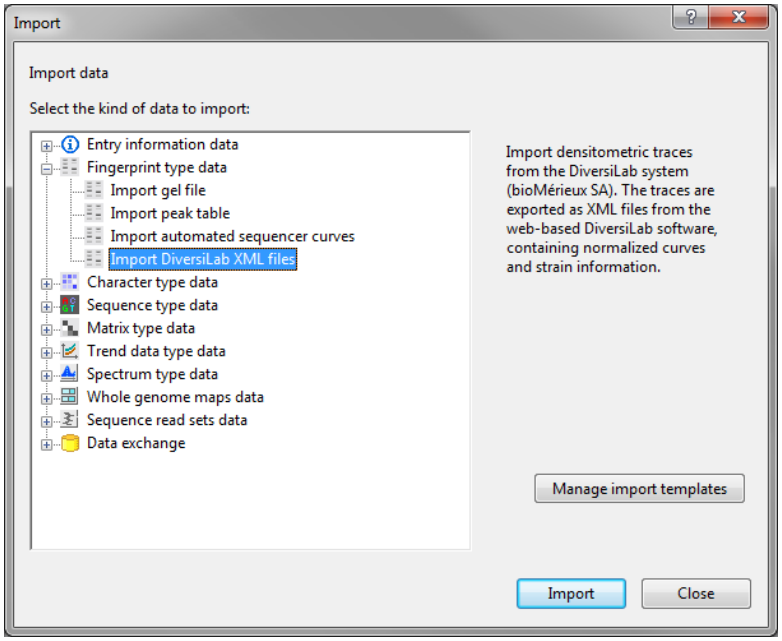

Figure 1.5: The ***Import DiversiLab XML files*** item in the Import tree.

Chapter 2

DiversiLab data analysis

2.1 Importing DiversiLab data

Each imported XML report file is saved as a single file, corresponding to one run or gel file in BioNumerics.

1.1 Select ***File*** > ***Import...*** ( , **Ctrl+I**) in the *Main* window to call the *Import* dialog box.

1.2 Select ***Import DiversiLab XML files*** from the Import tree under ***Fingerprint type data*** and press <***Import***>.

A first dialog box asks to select the files to import.

1.3 Browse for the DiversiLab XML files.

The second dialog box asks for the fingerprint type the gel files should be saved in (see Figure 2.1).

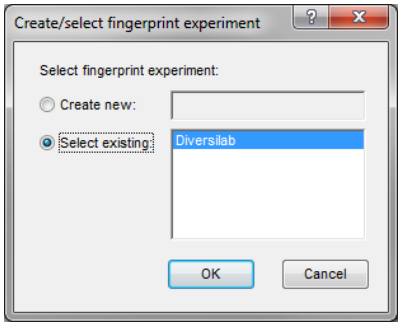

Figure 2.1: The *Create/select fingerprint experiment* dialog box.

You can select an existing fingerprint type for the storage of the gel files or create a new one. When the option ***Create new*** is selected, an experiment name needs to be specified in the text box before going to the next step of the wizard.

If a new fingerprint experiment type needs to be created, the following settings are asked for (see Figure 2.2):

- ***OD range*** (default 256): The ***OD range*** is the number of intensity levels the curves consist of (dynamic range). This number is sometimes also indicated as a bit depth; 8-bit corresponds to 256, 12-bit to 4096, and 16-bit to 65536 intensity levels. Other values can also be entered.
- ***Normalized track resolution*** (default 1000): The ***Normalized track resolution*** defines the resolution (track length, expressed in points) the traces will be rescaled to after normalization. A number should be entered that is equal to or less than the resolution of the imported traces.

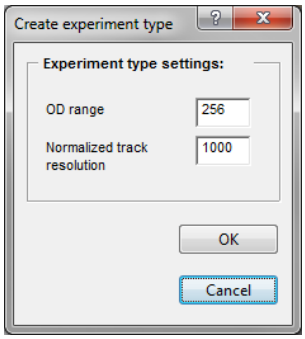

Figure 2.2: The *Create experiment type* dialog box.

1.4 Press the <***OK***> button to import the data in the database.

When the import is finished, a gel file is created with the report name plus the import date and time as filename (see the *Fingerprint files* panel). New database entries are created for the lanes, with automatically generated keys. The information fields provided in the DiversiLab XML files are created in the database if they did not exist yet, and the information for each entry is filled in (see Figure 2.3 for an example).

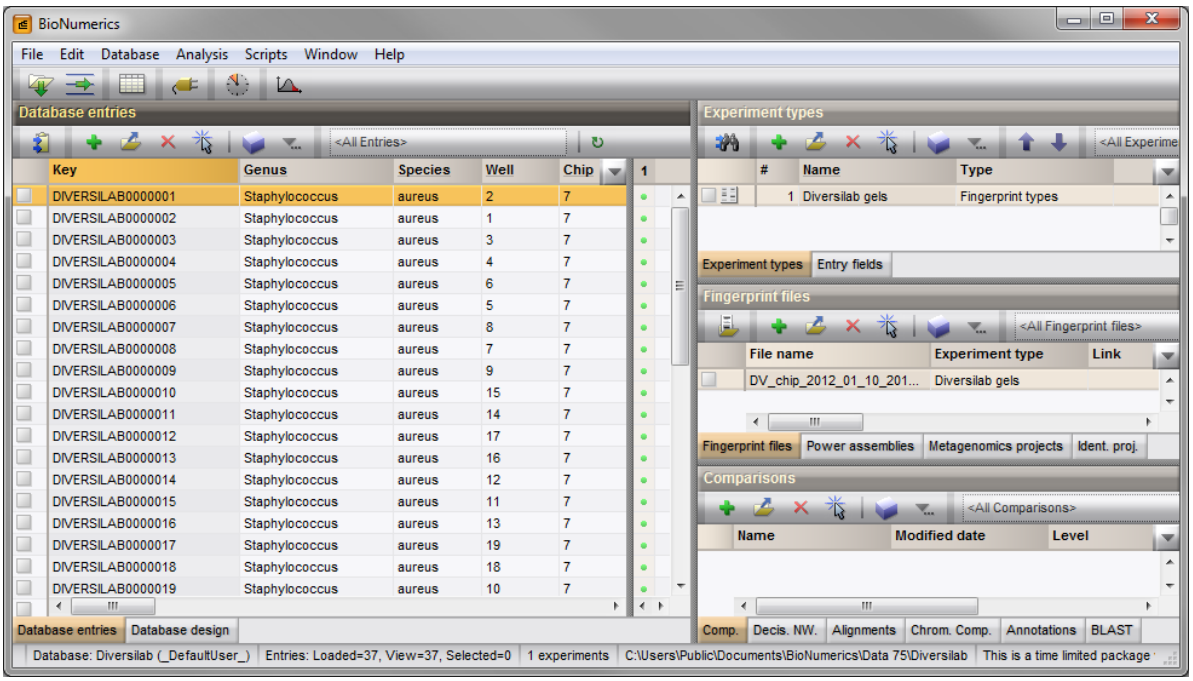

Figure 2.3: The *Main* window, after import of DiversiLab XML files.

Fingerprint file names cannot be longer than 80 characters and will be truncated. To avoid errors, it is advised to limit file names to maximum 60 characters.

2.2 Comparisons

2.2.1 Selections in BioNumerics

Before we can analyze the data, we need to make a selection in the database.

2.1 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.

2.2 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.

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

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

2.2.2 The Comparison window

2.5 Make a selection in the *Database entries* panel (see 2.2.1).

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

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

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

2.8 Display the densitometric curves as pseudo-gel strips with ***Layout*** > ***Show image*** ( ).

2.9 You may need to adjust the brightness/contrast using ***Fingerprints*** > ***Settings*** > ***Brightness & contrast...*** ( ).

2.10 Calculate a dendrogram with ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

2.11 Select ***Pearson correlation*** and enter 3% as ***Optimization*** in the *Similarity coefficient* wizard page, then press <***Next***> and select ***UPGMA*** in the *Cluster analysis* wizard page.

2.12 Press <***Finish***> to calculate the dendrogram.

Note that, although the traces are normalized between two external marker peaks (one at the top and one at the bottom of the traces), there is still a significant shift between the peak positions of very similar patterns. This shift is due to a phenomenon of random migration within wells, which is inherent to the Agilent microfluidics chips used in the DiversiLab system.

However, since we entered 3% as an optimization factor, most of the shifts are compensated for while calculating the correlation. As a result, the dendrogram exhibits clusters as if the traces were nicely aligned. Since the optimization is a pairwise correction, it does not result in a global alignment and as such, cannot be shown on the image. We refer to 2.3 for a global image alignment.

2.3 On-the-fly correction of curve alignments

2.3.1 Introduction

In order to correct for the normalization problems explained before, 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 2.4): 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

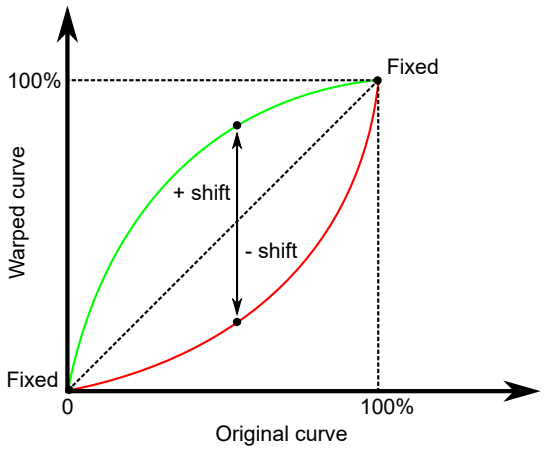

Figure 2.4: Non-linear shift with fixed edges.

shift is based on a quadratic function with one degree of freedom 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 2.5): 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 two degrees of freedom and is

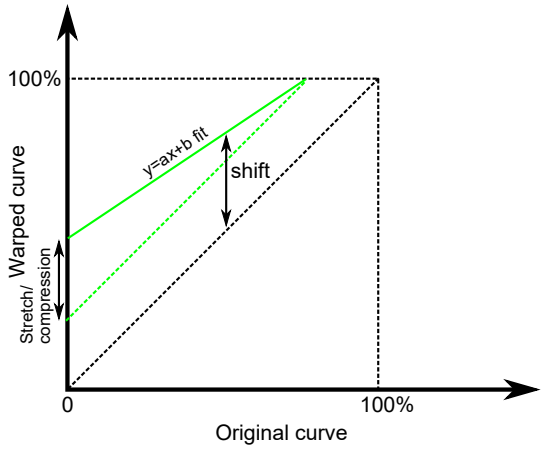

Figure 2.5: Global shift with a linear stretch or compression.

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.

2.3.2 Calculating the alignment

The auto-correction can only be done if a similarity matrix is present for the fingerprints (see 2.2). For best results, the ***Optimization*** value applied should be large enough, e.g. 3%. Furthermore, 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:

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

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

Four settings need to be provided (see Figure 2.6).

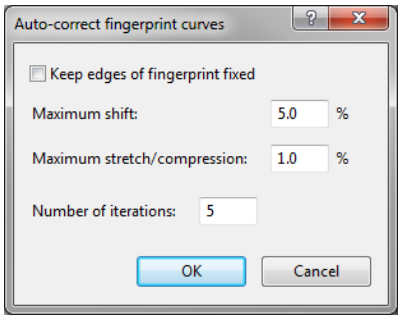

Figure 2.6: The *Auto-correct fingerprint curves* dialog box.

- ***Keep edges of fingerprints fixed***: if this setting is enabled, no global stretch or compression of the curve is possible (see Figure 2.4). If this setting is switched off, a linear stretch/compression is applied on the curve (see Figure 2.5).
- ***Maximum shift*** (as % of the curve length).
- ***Maximum stretch/compression*** (as % of the curve length).
- ***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 2.7).

3.3 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 2.8:

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:

3.4 In the *Main* window, double-click on the **DiversiLab** fingerprint type to open it.

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

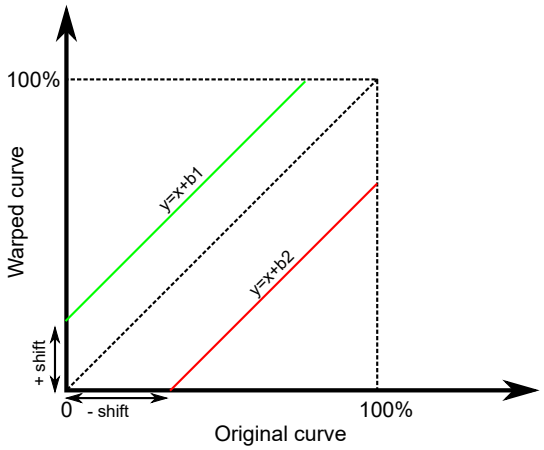

Figure 2.7: Linear stretch or compression applied, with a shift only.

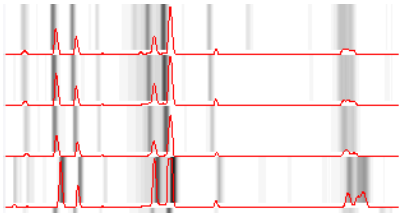

Figure 2.8: Detail

3.6 In the *Comparison* window, click inside the *Experiment data* panel to update the image. The corrected curves are now shown as gel strips.

3.7 You can repeat steps Instruction 3.2 to Instruction 3.3 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 2.9 illustrates the effect of the alignment without stretch/compression and with stretch/compression, respectively.

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.

3.8 A clustering can now be calculated as in 2.2 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.

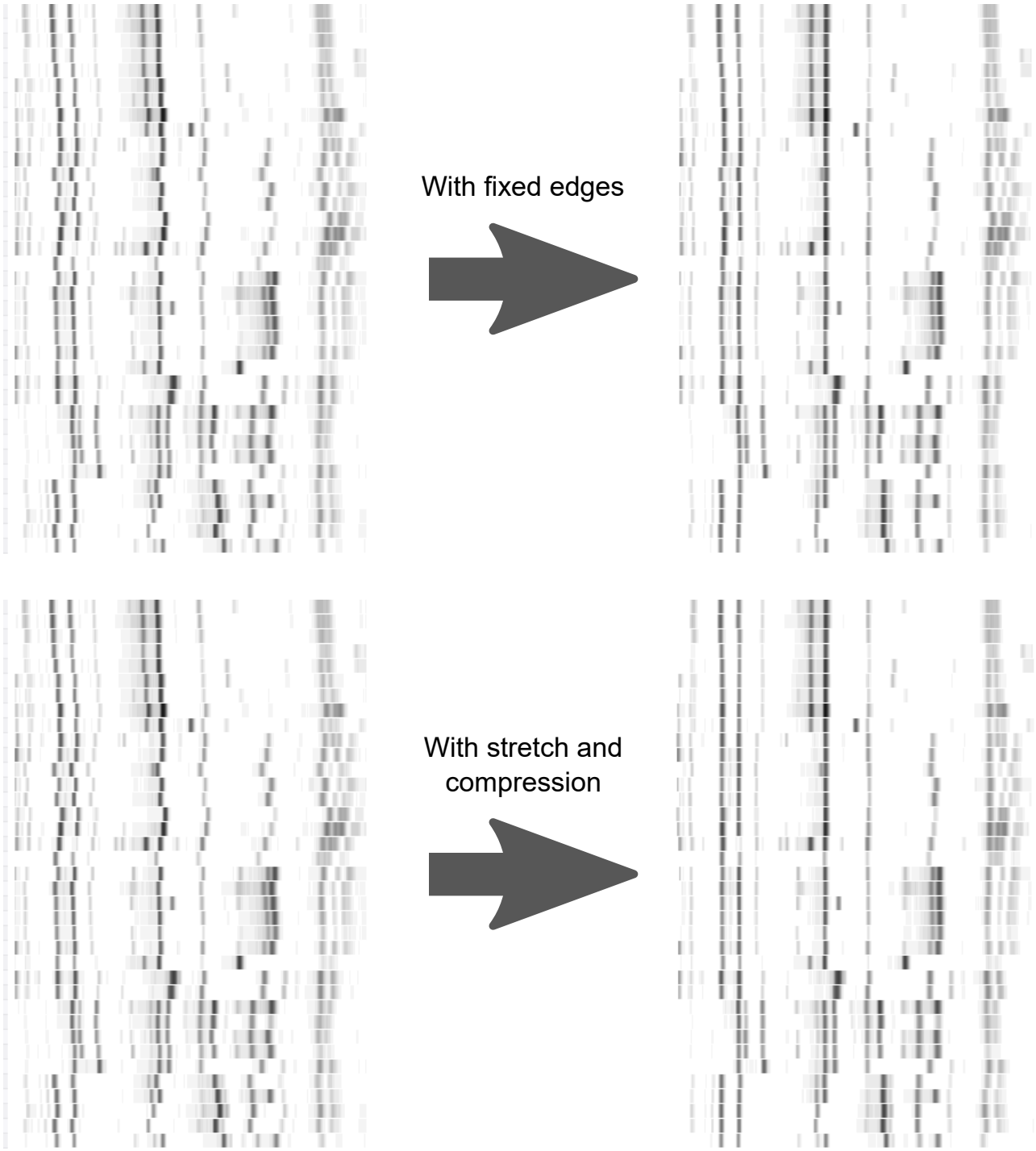

Figure 2.9: Effects of the different alignment options.

APPLIED MATHS

A BIOMÉRIEUX COMPANY

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

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

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