BioNumerics
Quick
Guide

BioNumerics Quick Guide
Version 6.6

APPLIED
MATHS

www.applied-maths.com

Contents

1 Database **3**

1.1 Getting started **5**
1.1.1 Introduction . . . 5
1.1.2 Software installation . . . 5
1.1.3 Quick Guide tutorial data . . . 9

1.2 Creating and setting up a new database **11**
1.2.1 Creating a new database . . . 11
1.2.2 Setting up a new database . . . 11
1.2.3 Selections of entries . . . 15
1.2.3.1 Manual selections . . . 15
1.2.3.2 Automatic search and select functions . . . 15

2 Experiments **17**

2.1 Fingerprint data **19**
2.1.1 Introduction . . . 19
2.1.2 Sample data . . . 19
2.1.3 Create a fingerprint type experiment . . . 19
2.1.4 Import a fingerprint gel image file . . . 20
2.1.5 Process the fingerprint gel file . . . 20
2.1.5.1 Define strips . . . 21
2.1.5.2 Define curves . . . 23
2.1.5.3 Normalize the gel . . . 24
2.1.5.4 Define bands . . . 25
2.1.6 Link fingerprint data to entries . . . 28
2.1.7 Fingerprint type experiment settings . . . 29
2.1.7.1 Assigning a standard pattern . . . 29
2.1.7.2 Calculating a calibration curve . . . 30
2.1.8 Additional practice . . . 31
2.1.8.1 XbaI-002 . . . 31
2.1.8.2 AvrII-001 . . . 32
2.1.8.3 AvrII-002 . . . 33

2.2 Character data **35**
2.2.1 Introduction . . . 35
2.2.2 Sample data . . . 35
2.2.3 Creating a character type experiment . . . 36
2.2.4 Importing character data from external files . . . 37
2.2.4.1 Importing character data from an Excel file . . . 37
2.2.4.2 Importing character data from a text file . . . 39
2.2.5 Changing the character type settings . . . 41

2.2.5.1 Experiment card settings . . . 41
2.2.5.2 Character color setup . . . 42

2.3 Sequence data **43**
2.3.1 Introduction . . . 43
2.3.2 Creating a sequence type experiment . . . 43
2.3.3 Importing FASTA sequences . . . 44
2.3.4 Importing chromatogram trace files . . . 46

3 Comparisons **51**

3.1 General comparison functions **53**
3.1.1 Comparison settings . . . 53
3.1.2 Comparing two entries . . . 53
3.1.3 Creating a new comparison . . . 54
3.1.4 Comparison window . . . 55
3.1.4.1 Comparison layout . . . 55
3.1.4.2 Add and remove entries . . . 56
3.1.4.3 Create groups . . . 56

3.2 Clustering fingerprint data **59**
3.2.1 Comparison window . . . 59
3.2.2 Clustering fingerprint data . . . 59
3.2.3 Matrix display functions . . . 62
3.2.4 Printing a cluster analysis . . . 62
3.2.5 Additional practice . . . 63

3.3 Clustering character data **65**
3.3.1 Comparison window . . . 65
3.3.2 Clustering character data . . . 66

3.4 Sequence alignment and clustering **69**
3.4.1 Introduction . . . 69
3.4.2 Comparison window . . . 69
3.4.2.1 Pairwise sequence cluster analysis . . . 69
3.4.2.2 Multiple sequence alignment . . . 70
3.4.2.3 Sequence cluster analysis based on multiple alignment . . . 72
3.4.2.4 Exporting a multiple alignment . . . 73
3.4.3 Alignment window . . . 73
3.4.3.1 Sequence display . . . 73
3.4.3.2 Alignment and clustering . . . 73
3.4.3.3 Mutation and SNP analysis . . . 74

3.5 Band matching tables **75**
3.5.1 Creating a band matching table . . . 75
3.5.1.1 Creating a composite data set . . . 75
3.5.1.2 Creating band classes . . . 76
3.5.1.3 Displaying the band matching table . . . 76
3.5.2 Band polymorphism analysis . . . 77
3.5.2.1 Finding discriminative band classes . . . 78
3.5.2.2 Sorting entries by band intensity . . . 78
3.5.3 Additional practice . . . 78

3.6 Composite data sets **79**
3.6.1 Introduction . . . 79
3.6.2 Combining character experiments . . . 79
3.6.2.1 Creating a composite character set . . . 79
3.6.2.2 Cluster analysis of a composite character set . . . 80
3.6.3 Combining fingerprint experiments . . . 81
3.6.3.1 Creating a composite data set . . . 81
3.6.3.2 Cluster analysis of a composite data set . . . 82
3.6.4 Additional practice . . . 83

3.7 Dimensioning techniques **85**
3.7.1 Multidimensional scaling (MDS) . . . 85
3.7.1.1 Calculating an MDS . . . 85
3.7.1.2 Changing the coordinate space layout . . . 86
3.7.2 Principal components analysis (PCA) . . . 87
3.7.2.1 Calculating a PCA . . . 87
3.7.2.2 Changing the PCA layout . . . 87

4 Identification **89**

4.1 Identification of unknown entries **91**
4.1.1 Identification of unknown entries in a comparison . . . 91
4.1.2 Identification of unknown entries in a library . . . 91
4.1.2.1 Creating a library . . . 92
4.1.2.2 Identifying unknown entries with a library . . . 93

4.2 Decision networks **95**
4.2.1 Introduction . . . 95
4.2.2 Creating a new decision network . . . 95
4.2.3 Building a decision network . . . 96

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.
13809 Research Boulevard, Suite 645
Austin, Texas 78750
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.

Part 1

Database

Chapter 1.1

Getting started

1.1.1 Introduction

This Quick Guide provides a general introduction to BioNumerics. Since BioNumerics is a powerful and complex application, many features are not covered in the Quick Guide. Please refer to the Manual for more detailed information on these features. A PDF version of the manual can be found in the installation directory of BioNumerics.

While it is not necessary to follow the entire Quick Guide, Chapter 1.1 and Chapter 1.2 are essential for all new users. In these Chapters, you will learn how to install the software, create a new database, and add new database entries. You will then be prepared to import data in the Chapters covering fingerprints (Chapter 2.1), characters (Chapter 2.2), and sequences (Chapter 2.3). If you prefer to start with data analysis, you can directly go to the *Comparisons* and *Identification* Parts (Part 3 and Part 4, respectively).

Your ability to use each Chapter will depend on which modules are included with your license. For example, the *Identification* Chapters require the Identification module; the *Dimensioning techniques* Chapter requires the Dimensioning and Statistics module, and so on. All modules are included when evaluating the software.

1.1.2 Software installation

If you have not already installed BioNumerics, locate the CD-ROM that came with the package. The latest version of the software can also be downloaded from the Applied Maths website: go to `http://www.applied-maths.com`, select *Download* and *Software*.

2.1 Launch the Setup executable.

During a first-time installation of BioNumerics, the *Welcome dialog box* will display the version number of BioNumerics that is included with the Setup package (see Figure 1.1.1).

2.2 Please verify that you are installing the correct version and click <***Next***> to continue.

If an instance of BioNumerics 6.1 or older is already installed, the update *Welcome dialog box* will be displayed. This dialog box shows the version number of the installed instance of BioNumerics and the new version. The wizard will offer the choice between the installation of a new BioNumerics instance (choose a new installation directory) or to upgrade the existing instance (choose same installation directory as older version).

If an instance of BioNumerics 6.5 is already installed, the *Existing Installed Instances Detected dialog box* will appear when launching the Setup executable. This dialog box allows you to choose between installing a new BioNumerics instance or changing an existing instance.

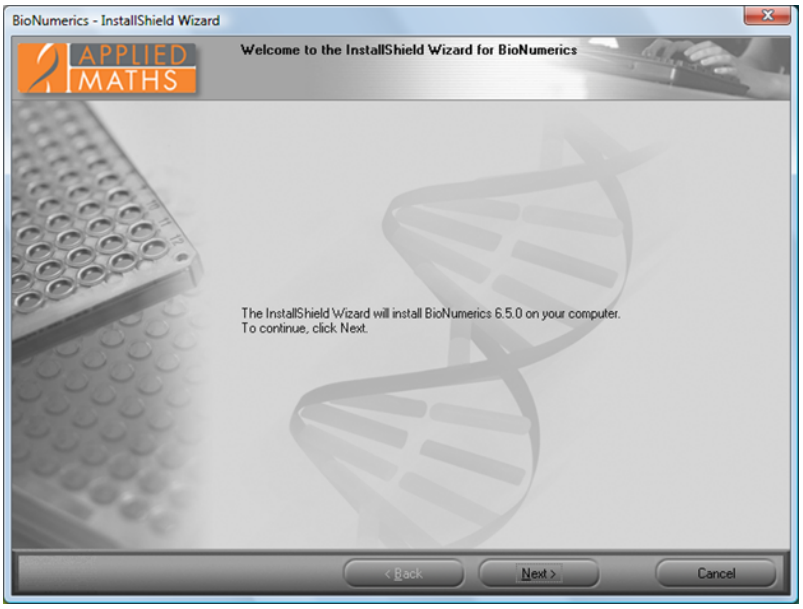


Figure 1.1.1: The *Welcome dialog box*.

The next dialog will display the Software End User License Agreement (EULA).

2.3 Please read the EULA carefully and click the top *I accept the terms of the license agreement* radio button and the <***Next***> button to continue the installation.

The user name, organization name and BioNumerics license string need to be entered in the *Customer Information dialog box* (see Figure 1.1.2). The license string is provided on the sleeve of the CD–ROM or in case of an upgrade or an internet evaluation license, you may have obtained it electronically.

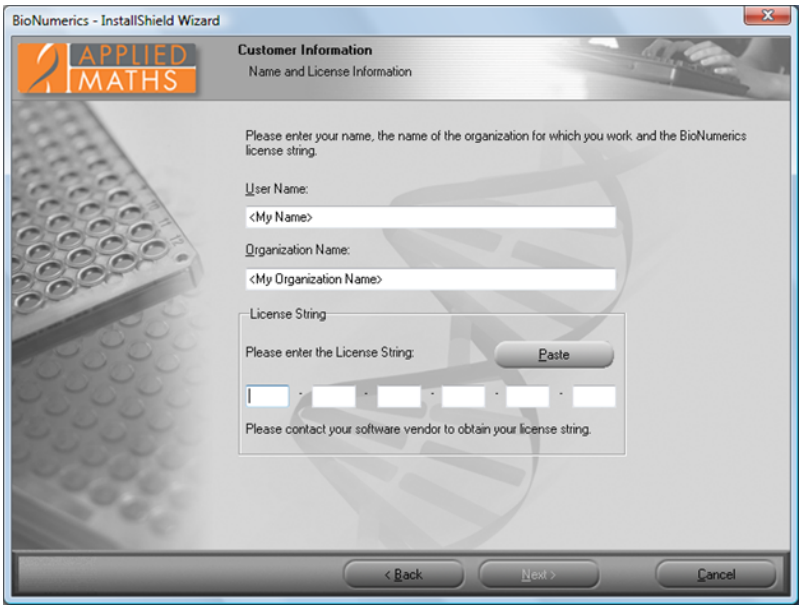


Figure 1.1.2: The *Customer Information dialog box*.

2.4 Specify the user and organization names, enter the license string and press <***Next***>.

You must enter a valid license string to be able to continue with the installation. In addition the user and organization names cannot be empty.

The installation directory for the BioNumerics application and the database home directory can be entered in the *Choose Destination Location dialog box* (see Figure 1.1.3).

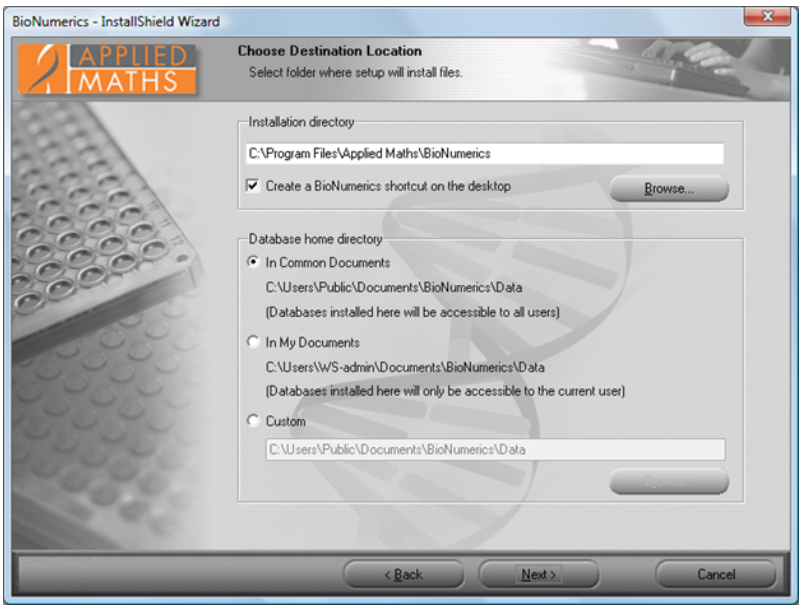


Figure 1.1.3: The *Choose Destination Location dialog box*.

2.5 Make sure the correct folders are specified and press <***Next***> once more.

The BioNumerics features that you want to install on the local computer can be selected in the *Select Features dialog box* (see Figure 1.1.4). Clicking on a feature in the left pane will display a short description in the right pane.

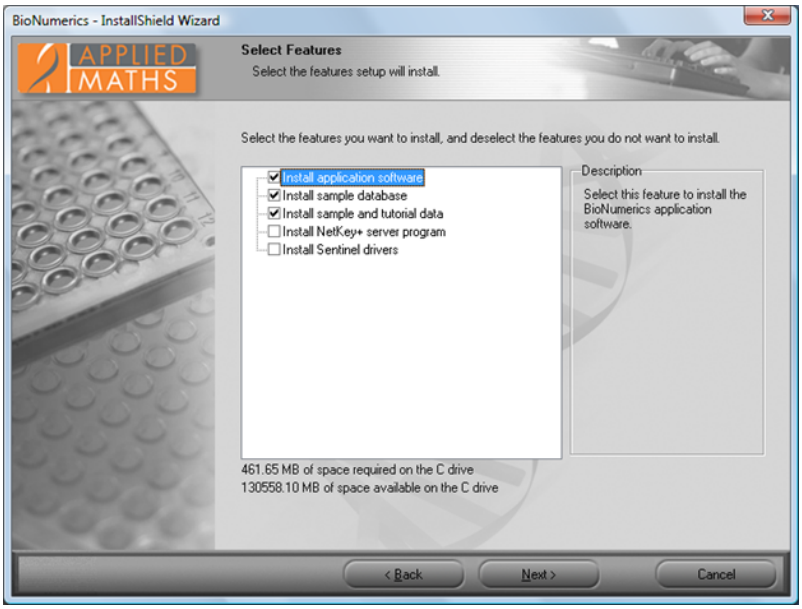


Figure 1.1.4: The *Select Features dialog box*.

Install application software:

- In case of a *standalone license*, the *Application software* needs to be installed on each computer that you want to use to run the software. Please note that only on the computer where the dongle is attached to, you will be able to work with the software.

- In case of an *internet license*, the *Application software* needs to be installed on the computer that you want to use to run the software. Please note that a permanent and stable internet connection is required to run the internet license.
- In case of a *network license*, the *Application software* needs to be installed on the computers in the network that you want to use to run the software.

Install sample database and Install sample and tutorial data:

- The *Sample database* and *Sample and tutorial data* that are contained in the Setup package are used in the Quick Guide and in the Manual to illustrate the features of the software. Selecting these features will install the *Sample database* and *Sample and tutorial data* in the BioNumerics home directory that is specified in the *Choose Destination Location dialog box* (see Figure 1.1.3).

Install Sentinel drivers:

- In case of a *standalone license*, the *Sentinel drivers* need to be installed on each computer that you want to use to run the software.
- In case of an *internet license*, you only need an internet connection to run the software. Since no USB dongle is needed to run an internet license, the *Install Sentinel drivers* option does not need to be checked.
- In case of a *network license*, the *Sentinel drivers* only need to be installed on the NetKey+ server computer in the network where the hardware security key will be connected to.

Install NetKey+ server program:

- The *NetKey+ server program* feature will only be visible and available for installation if a network license string has been entered in the *Customer Information dialog box* (see Figure 1.1.2). The *NetKey+ server program* feature must only be installed on the computer in the network where the hardware security key will be connected to.

2.6 Tick the appropriate check boxes for the features you want to install and press <***Next***>.

If a network license string was entered in the *Customer Information dialog box* (see Figure 1.1.2), and if the BioNumerics application feature was selected for installation (see Figure 1.1.4), the *NetKey+ connection settings dialog box* will pop up where the *NetKey+ Server name* and *Server port number* connection parameters can be entered.

2.7 Click <***Install***> to start the installation.

The *Setup Status dialog box* is displayed.

If a network license string has been entered in the *Customer Information dialog box* (see Figure 1.1.2), and the *NetKey+ server program* feature was selected for installation (see Figure 1.1.4), the Setup will ask if you want to run the NetKey+ Configuration tool. This tool allows you to install and subsequently start the NetKey+ service.

2.8 Press <***Finish***> to close the *InstallShield Wizard*.

2.9 Double-click on the shortcut on your desktop to open BioNumerics. Alternatively, open the start menu and select BioNumerics under *All programs*.

The Startup program appears (see Figure 1.1.5).

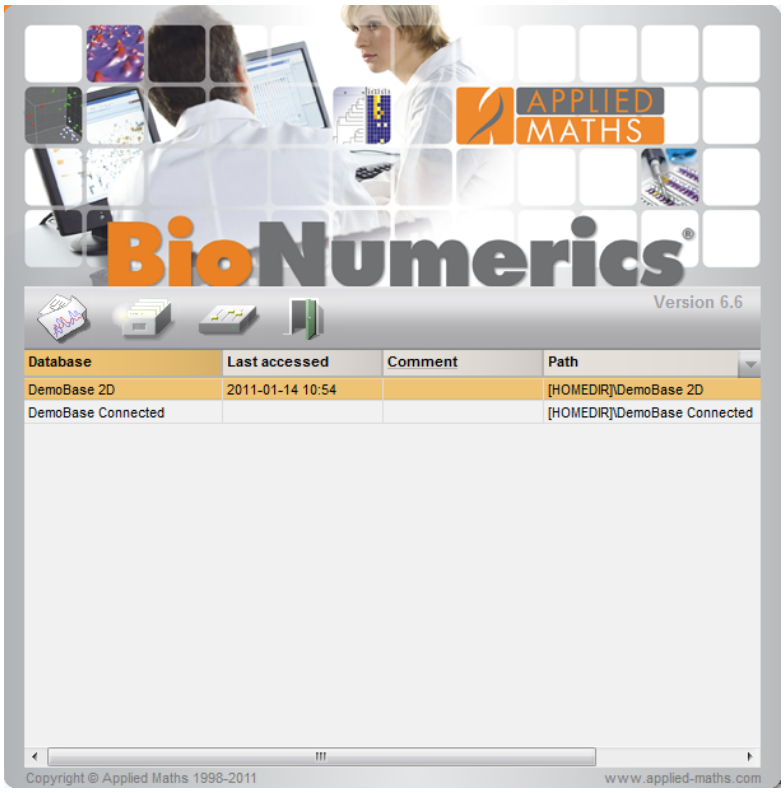


Figure 1.1.5: The BioNumerics Startup screen.

1.1.3 Quick Guide tutorial data

The data used in the first two Parts of the Quick Guide (*Database* and *Experiments*) can be installed with the software (check *Install sample and tutorial data* in the *Select Features dialog box*, see Figure 1.1.4). Alternatively, the data can be downloaded from the Applied Maths website: go to `http://www.applied-maths.com`, select *Download*, *Sample data*, and *BioNumerics Tutorial Data*.

The *Comparisons* and *Identification* Parts use the **DemoBase Connected** database. This database can be installed with the application software (check *Install sample database* in the *Select Features dialog box*, see Figure 1.1.4).

Chapter 1.2

Creating and setting up a new database

1.2.1 Creating a new database

A BioNumerics database is a collection of samples which are potentially comparable to each other. Although a single database can be organized into subsets, very large databases can be unwieldy, so it is preferable to maintain unrelated samples in separate databases. As an example, we will create a new database containing closely related *E. coli* samples.

1.1 In the BioNumerics Startup screen (see Figure 1.1.5), press the  button to enter the *New database wizard*.

1.2 Specify **E. coli** as the new database name and press <***Next***> twice.

1.3 Press <***Finish***> and press <***Proceed***> to set up the new database as a connected Access database.

1.4 Press <***Proceed***> in the *Plugin installation toolbox*. Plugins will be installed later.

A new, empty database opens (see Figure 1.2.1).

1.2.2 Setting up a new database

As an exercise, we will import data from the text file `Ecoli-info.txt` (see Figure 1.2.2) into our **E. coli** database. If the *Install sample and tutorial data* feature was checked in the *Install wizard* (see Figure 1.1.4), this text file can be found in the `Sample and Tutorial data` folder in the database home directory. Alternatively, this text file can be downloaded from our website: go to `http://www.applied-maths.com`, select *Download*, *Sample data*, and *BioNumerics Tutorial Data*.

2.1 Select *File* > *Import* or press  to call the Import tree.

2.2 In the Import tree, expand *Entry information data*, highlight *Import fields (text file)* and press <***Import***>.

2.3 In the *Select file* step, browse for the `Ecoli-info.txt` file in the `BioNumerics Tutorial data \Database` folder. Select the file, select the *TAB* separator from the *Field separator* list and press <***Next***> (see Figure 1.2.3).

The way the entry information should be imported from the selected file into the database needs to be specified with an import template.

2.4 Press the <***Create new***> button to create a new import template.

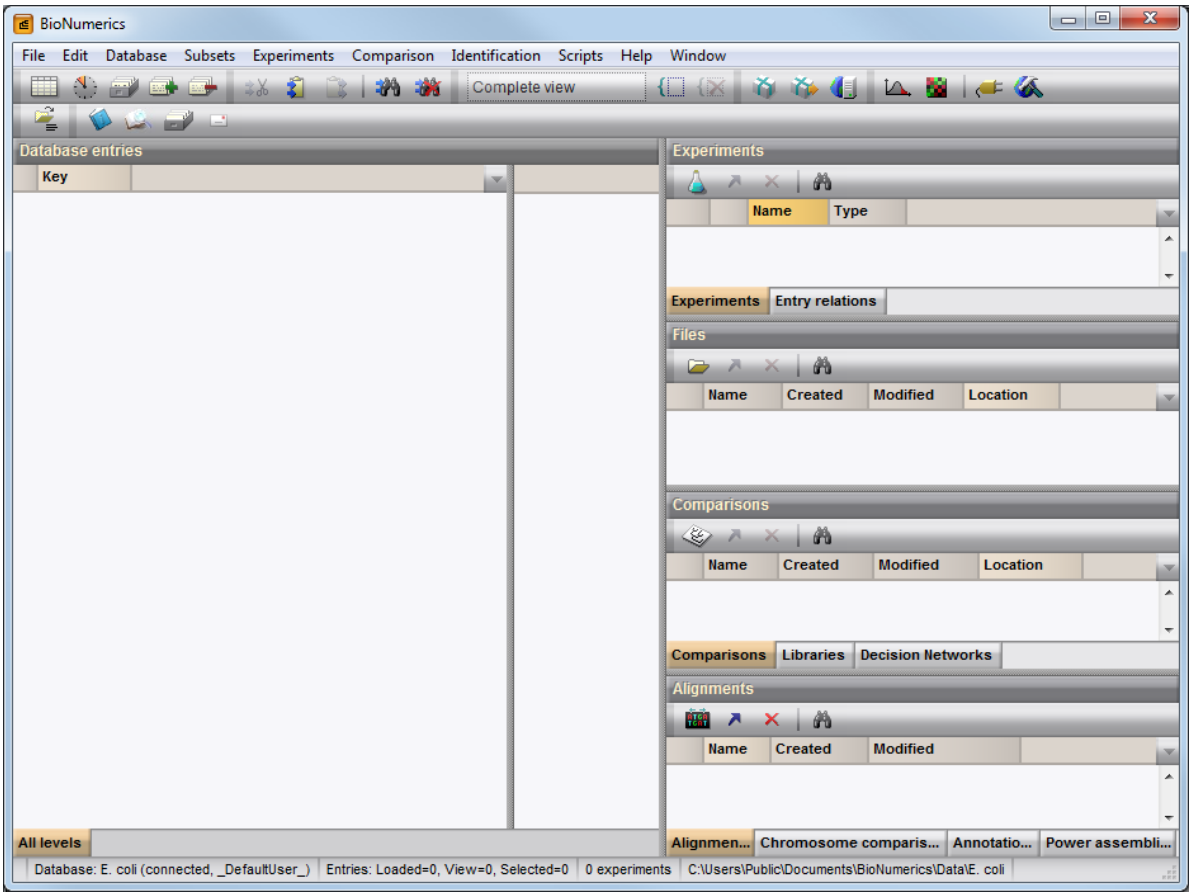


Figure 1.2.1: The *BioNumerics main window*.

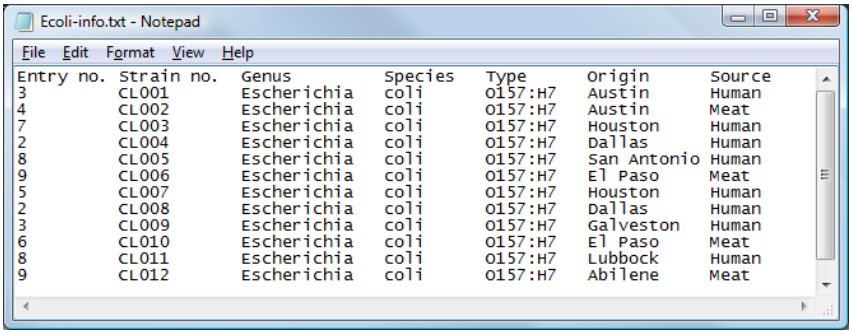


Figure 1.2.2: Import entries from a text file.

This brings up a new dialog box (see Figure 1.2.4). Each column in the selected file corresponds to a row in the grid (column 1 in the file corresponds to row 1 in the grid, column 2 corresponds to row 2, etc.). The text *File field* is specified in the *Source type* column and the column names are displayed in the *Source* column. The last row in the grid holds the name of the file.

2.5 Select the second row in the grid, press the <***Edit destination***> button and select the BioNumerics *Key* field from the list. Press <***OK***>.

The grid is updated (see Figure 1.2.4).

2.6 Highlight the five other external fields in the grid panel using the Shift-key (make sure the last row in the grid is not selected). Press the <***Edit destination***> button and select the *Entry info field* option from the list. Press <***OK***>.

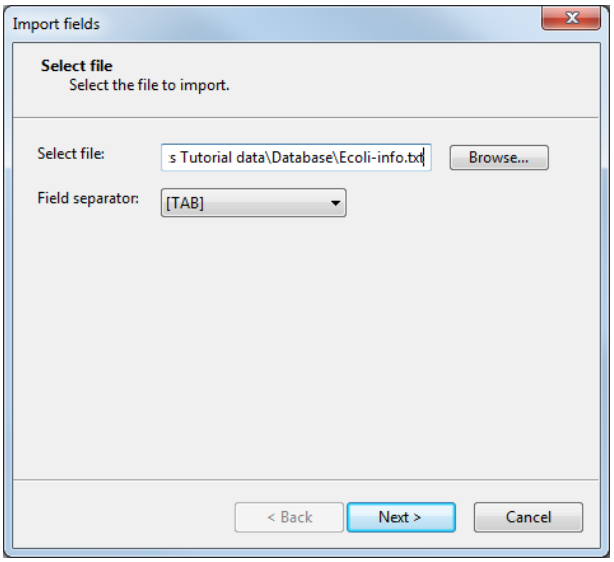


Figure 1.2.3: Select the file to import and the separator.

2.7 Press <***OK***> once more to accept the default suggested names and press <***Yes***>.

The grid is updated (see Figure 1.2.4).

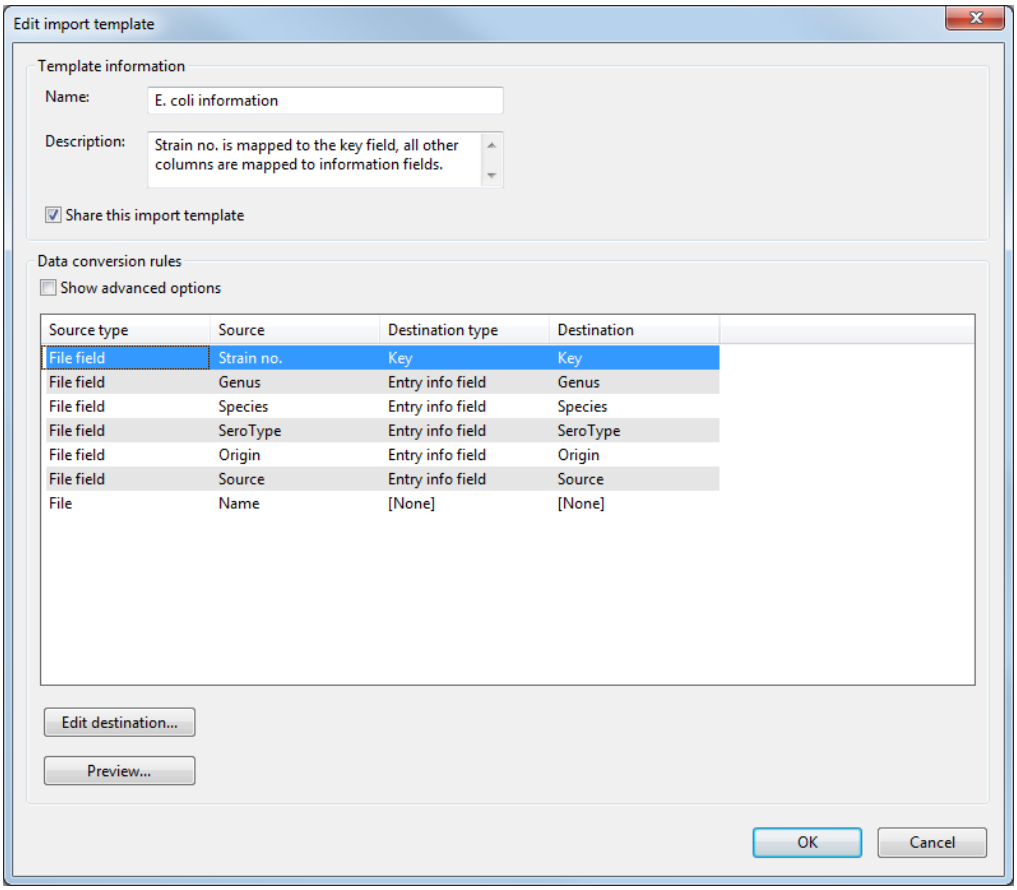


Figure 1.2.4: Define a new import template.

2.8 Optionally, change the default suggested template *Name* and press <***OK***>.

The import template is added to the list and is automatically selected.

2.9 Press <***Next***> twice.

The **E. coli** database now contains twelve new entries, each with six database information fields filled in (see Figure 1.2.5).

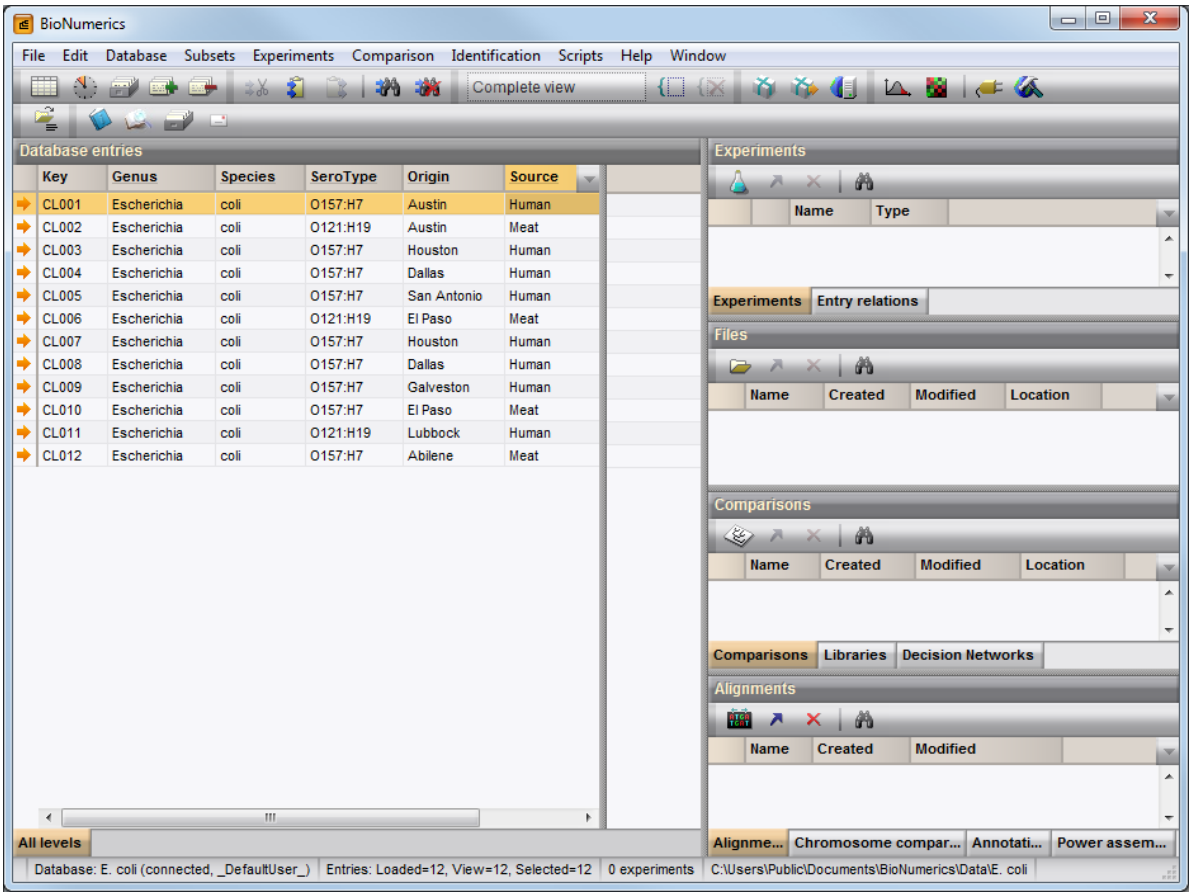


Figure 1.2.5: The *BioNumerics main window* after import.

2.10 Double-click on database entry **CL001** to open the *Entry edit window*.

Information can be edited in the *Entry edit window* and saved (see Figure 1.2.6).

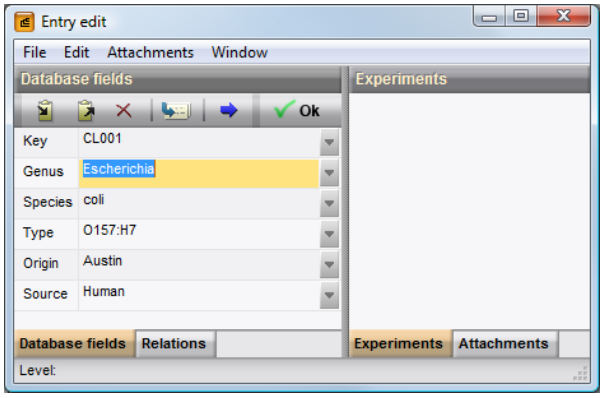


Figure 1.2.6: The *Entry edit window*.

2.11 Close the *Entry edit window*.

Alternative to using the *Entry edit window*, information in the information fields can be edited directly by clicking twice (not double-click) on an information field in the database. The information will appear highlighted and can be edited.

2.12 Click twice on one of the information fields (not a *Key* information field) or select **Ctrl+Enter**.

The information appears selected blue against a bright colored background and can be modified (see Figure 1.2.7).

Database entries

Key	Genus	Species	Serotype
CL001	Escherichia	coli	O157:H7
CL002	Escherichia	coli	O157:H7
CL003	Escherichia	coli	O157:H7
CL004	Escherichia	coli	O157:H7
CL005	Escherichia	coli	O157:H7

Figure 1.2.7: Clicking twice on an information field enables direct editing.

2.13 Use the **ArrowUp** and **ArrowDown**–keys on the keyboard to jump to the previous/next row.

2.14 To jump to the next column, use the **Tab**–key.

2.15 To jump to the previous column, select **Shift+Tab** on the keyboard.

1.2.3 Selections of entries

1.2.3.1 Manual selections

3.1 To select an entry in the database, hold the **Ctrl**–key and click on the entry in the *Database entries panel*. Alternatively, use the space bar to select entries.

The selected entry is marked by a colored arrow (see Figure 1.2.8).

Database entries

Key	Genus	Species	Serotype
CL001	Escherichia	coli	O157:H7
CL002	Escherichia	coli	O157:H7
CL003	Escherichia	coli	O157:H7
CL004	Escherichia	coli	O157:H7

Figure 1.2.8: A selected entry.

3.2 Selected entries are unselected in the same way (**Ctrl+click**).

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

All entries that are listed between these two entries are selected, including the two entries.

3.4 To select all entries in the database, use *Edit* > *Select all entries* (**Ctrl+A**).

3.5 To clear a selection in the database, use *Edit* > *Unselect all entries* (, **F4**).

1.2.3.2 Automatic search and select functions

Simple and intuitive search functions can be used to search and select entries.

3.6 Select *Edit* > *Search entries...* (, **F3**).

3.7 In the *Entry search window*, enter for example ”Dallas” in the *Origin* field (see Figure 1.2.9) and press <***Search***>.

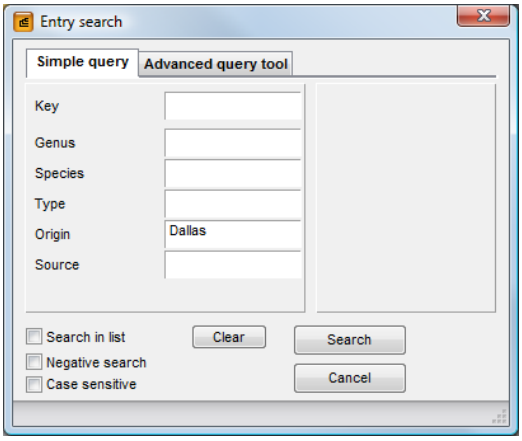


Figure 1.2.9: The *Entry search window*.

All entries having the name ”Dallas” in their ’Origin’ field are selected.

3.8 To clear the selection, use *Edit* > *Unselect all entries* (, **F4**).

Part 2

Experiments

Chapter 2.1

Fingerprint data

2.1.1 Introduction

In this Chapter we will:

- Create a fingerprint type experiment
- Import a fingerprint gel image file
- Process the fingerprint gel file
- Link fingerprint data to entries

2.1.2 Sample data

As an exercise, we will import an 8–bit TIFF gel image which was generated by PFGE with the restriction enzyme **Xba-I** in our **E. coli** database. The gel image can be downloaded from the Applied Maths website: go to `http://www.applied-maths.com/download/sampledata.htm` and click on "BioNumerics Tutorial Data". If the *Install sample and tutorial data* feature was checked in the *Install wizard* (see Figure 1.1.4), the gel image can be found in the `Sample and Tutorial data` folder in the database home directory.

2.1.3 Create a fingerprint type experiment

First we need to create a new fingerprint type experiment before we can import a gel image file in our **E. coli** database.

3.1 In the BioNumerics startup screen, double-click on the **E. coli** database – created in 1.2.1 – to open it.

3.2 In the *BioNumerics main window*, select *Experiments* > *Create new fingerprint type...* or press the button from the *Experiments panel toolbar* and select *New fingerprint type*.

3.3 Enter a name, for example **PFGE-XbaI** and press <***Next***>.

3.4 In the next dialog box, select *Two–dimensional TIFF files* and *8–bit OD depth (256 gray values)*. Press <***Next***> to proceed.

3.5 In the next dialog box, select *Yes* for fingerprints with inverted densitometric values.

3.6 Press <***Next***> to proceed.

3.7 In the final step, leave *No* selected for applying a background subtraction.

3.8 Press <***Finish***>.

The *Experiments panel* now lists the fingerprint type **PFGE–XbaI** (see Figure 2.1.1).

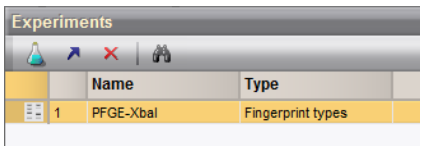


Figure 2.1.1: The *Experiments panel*.

You can still change the resolution, inversion, and background subtraction settings later when processing a gel.

2.1.4 Import a fingerprint gel image file

4.1 Select *File* > *Add new experiment file...* ( ) in the *BioNumerics main window*.

4.2 Select the file `ec-XbaI-001.tif` in the `BioNumerics Tutorial data\PFGE TIFFS` folder.

A box appears asking if you want to edit the image. Press <***No***> if you are sure the file is an uncompressed gray scale TIFF image. For the conversion to an uncompressed gray scale TIFF file press <***Yes***>.

4.3 Since the example file is an uncompressed gray scale TIFF file, press <***No***>.

The gel image is now available in the *Files panel* (see Figure 2.1.2). The file is marked with a red "N" ( ) indicating that the image has not been processed yet.

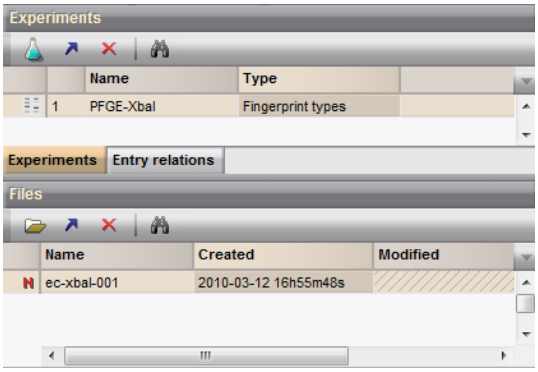


Figure 2.1.2: The *Experiments panel* and *Files panel*.

2.1.5 Process the fingerprint gel file

Before linking our fingerprint lanes to entries in our database, we must define and process the lanes in the *Fingerprint data window*.

5.1 In the *Files panel*, double–click on the filename **ec-XbaI-001** to open the *Fingerprint file window*.

5.2 In the *Fingerprint file window*, select *File* > *Edit fingerprint data* to open the *Fingerprint data window*.

5.3 In the next dialog box, select **PFGE-XbaI** and press <***OK***>.

The *Fingerprint data window* opens and the imported gel image is surrounded by a green rectangle. Since BioNumerics recognizes the darkness as the intensity of a band, make sure the bands appear as dark bands on a white background (see Figure 2.1.3).

In case the bands appear as white bands on a black background, invert the values: press to open the *Fingerprint conversion settings window*, check/uncheck the *Inverted values* check box, and press <***OK***> to apply the changes.

2.1.5.1 Define strips

The first step in processing a gel is to crop the image to remove empty space, and to define the lanes.

5.4 Delineate the area of the gel lanes by clicking and dragging the nodes of the rectangle to adjust it. Exclude the wells from the rectangle (see Figure 2.1.3).

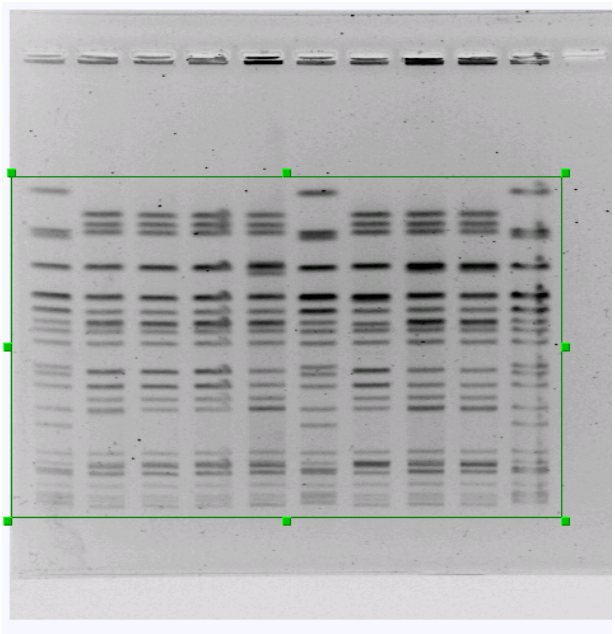


Figure 2.1.3: Area of gel lanes.

5.5 In the toolbar, press to let the software search for the individual lanes.

5.6 Enter "10" as the approximate number of lanes and press <***OK***>.

5.7 In the toolbar, press to open the *Fingerprint conversion settings window*.

When opened in the *Strips panel* the *Raw data tab* is selected by default.

5.8 Adjust the *Thickness* of the image strips so the splines border the edges of the bands (e.g. **33** points).

5.9 Press <***OK***>.

5.10 Adjust the horizontal position of each spline as necessary by clicking and dragging a blue node. Hold down the **Shift**–key while dragging a node to adjust the spline locally.

5.11 Decrease the thickness of the spline in lane **10** by clicking one of its nodes and repeatedly pressing or **F8**. Then move the spline over to the good portion of the lane to the left (see Figure 2.1.4).

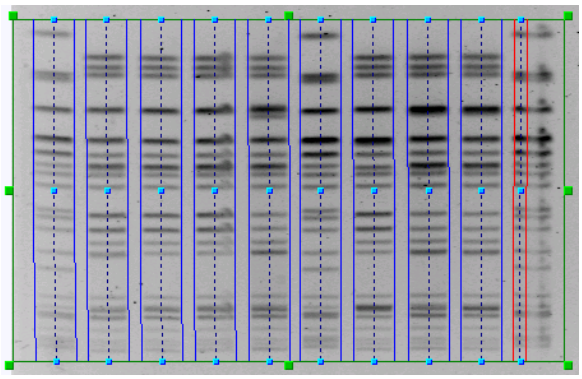


Figure 2.1.4: Adjusting the thickness and positions of the splines.

Next, we will edit the tone curve to improve the band visibility.

5.12 From the pull down menu, select *Edit* > *Edit tone curve*.

5.13 Press <***Linear***>. The visible gray scale interval now ranges between the minimum and maximum values in the image (see Figure 2.1.5).

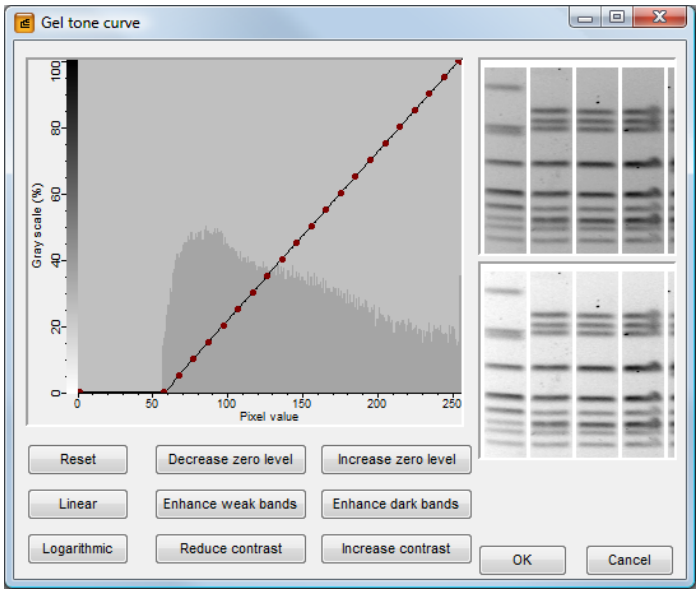


Figure 2.1.5: The *Gel tone curve window*.

5.14 Press the other editing buttons such as <***Enhance weak bands***> a few times until you have a clear and sharp image.

5.15 Press <***OK***> to save changes to the tone curve settings.

The gel tone curve does not change the image's densitometric values, only the way they appear in the *Fingerprint data window*.

5.16 Press to save the work already done.

5.17 Press to proceed to the **Curves** step or press the *Curves tab*.

2.1.5.2 Define curves

Now that the lanes have been defined, the software can generate densitometric curves describing the optical density across the spline along each lane. The left panel shows the strips extracted from the image file, the right panel shows the densitometric curve of the selected pattern (see Figure 2.1.6). The area between the blue lines of each lane will be used to calculate the densitometric curve.

5.18 Select lane **3**.

Near the top of the lane in the spline there is a pinpoint spot. As a result, a thin peak appears on the densitometric curve shown on the right side of the window.

5.19 It might help to zoom in on the image using the zoom slider.

5.20 Open the *Fingerprint conversion settings window* again (click on the button).

5.21 Change the *Averaging thickness* for curve extraction to **15** in the *Densitometric curves tab*.

5.22 Check *Median filter* and press <***OK***>.

In lane 3, the thin peak – resulting from the pinpoint spot – has disappeared in the densitometric curve shown on the right side of the window (see Figure 2.1.6).

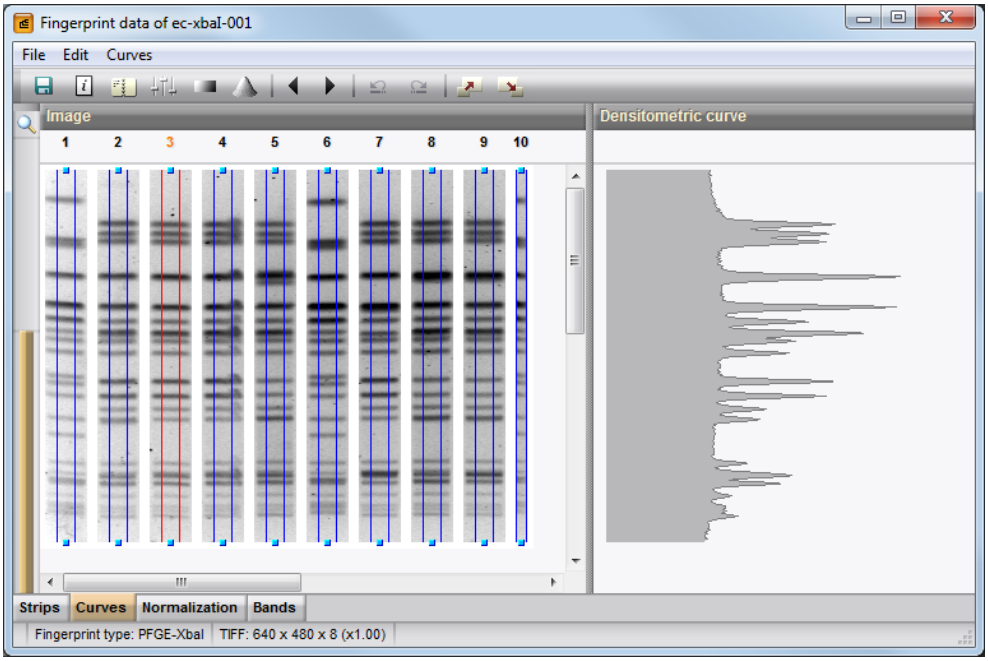


Figure 2.1.6: Increased splines and median filtering.

BioNumerics can analyze the gel image to determine the optimal settings for removing background noise from the densitometric curves.

5.23 Select *Curves* > *Spectral analysis*.

The following settings are shown in the *Spectral analysis window*:

1. **Wiener cutoff scale**: Determines the optimal setting for least square filtering.
2. **Background scale**: Estimation of the disk size for background subtraction.

5.24 Close the *Spectral analysis window*.

5.25 Open the *Fingerprint conversion settings window* again (click on the button).

5.26 Check *Apply least square filtering* and specify a *Least square filtering Cut off* as indicated by the Wiener cutoff scale in the *Spectral analysis window* (use the percentage value, e.g. **1**). Least square filtering removes very small peaks from the curves.

5.27 Check *Apply* in the *Background subtraction panel* and specify a *Background subtraction disk size* as indicated by the background scale in the *Spectral analysis window* (use the percentage value, e.g. **14**). Background subtraction removes large background trends from the curves.

5.28 Press <***OK***>.

The background noise has been removed from the curves (see Figure 2.1.7).

Figure 2.1.7: Curves after filtering.

5.29 Press to save the work already done.

5.30 Press to proceed to the **Normalization** step or press the *Normalization tab*.

2.1.5.3 Normalize the gel

Every fingerprint type experiment needs at least one reference system to normalize its gels. Since this is the first Xba-I gel we have imported, we need to create a reference system based on the standard pattern in this gel. We will use the standard's molecular weights to name the reference positions. Subsequent Xba-I gels, provided they contain the same standard, will be normalized with the same reference system.

5.31 Press to enter the normalized view.

For now, the "normalized view" looks the same as the original view. In a first step we will define the reference lanes and the reference bands.

5.32 Select lane **1** and press to assign it as a reference lane or select *References* > *Use as reference lane*.

5.33 Repeat this for lane **6** and lane **10**.

5.34 Choose the most suitable standard lane for creating the reference system. For this exercise select lane **6** by clicking anywhere in the lane so that the lane number turns orange.

5.35 Right-click on the top band in lane **6** and select *References* > *Add external reference position* from the pop-up menu.

5.36 Enter **582.6** and press <***OK***>.

5.37 Repeat the process for each band in lane **6** as shown in Figure 2.1.8.

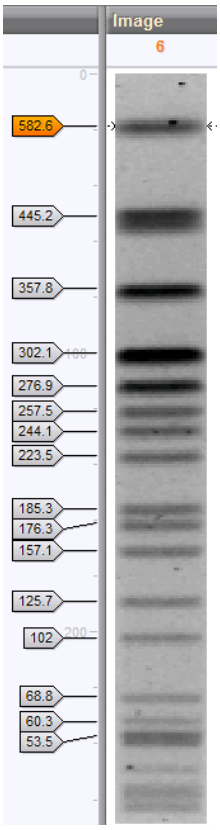


Figure 2.1.8: Reference system.

The reference system for experiment type **PFGE-XbaI** is now defined (see Figure 2.1.9).

5.38 Select *Normalization* > *Auto assign* or press .

5.39 Make sure *Using bands* is selected and press <***OK***>.

5.40 Carefully inspect the assignments made (see Figure 2.1.9).

5.41 If a band assignment is incorrect, select the band and press the **Del**–key.

5.42 To assign a reference band manually, first click on the reference position tag, then hold the **Ctrl**–key and click on the reference band to assign it to that reference position.

5.43 To update the normalization based on the band assignments, select *Normalization* > *Update normalization* or press .

5.44 Press to proceed to the last step. Alternatively press the *Bands tab*.

2.1.5.4 Define bands

If you want to use the curves to compare the patterns, no bands need to be assigned and this last step can be skipped. If you want to compare the patterns using bands, you will need to assign bands in the sample lanes

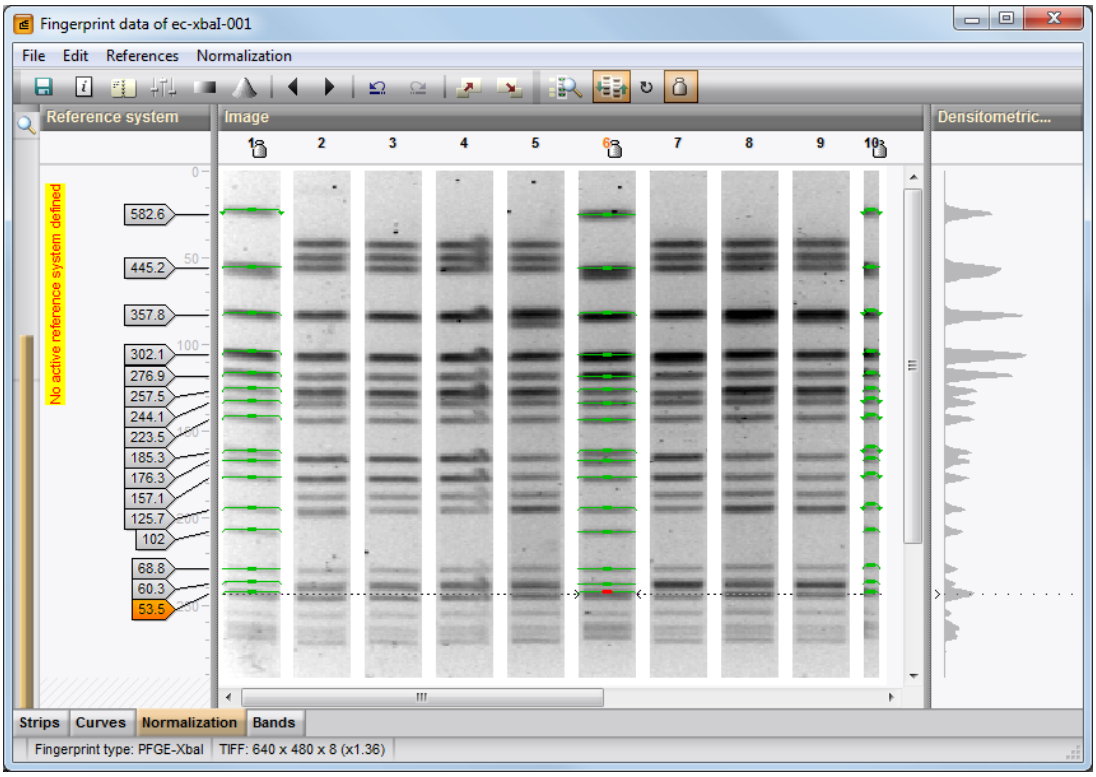


Figure 2.1.9: Reference positions assigned.

in this 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.

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

In the *Band search window*, the currently selected lane is shown along the bottom (see Figure 2.1.10).

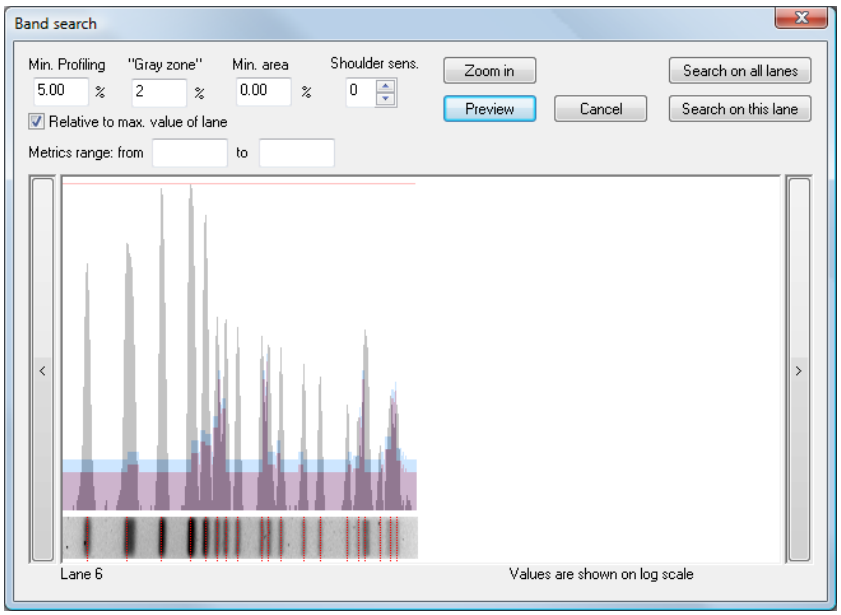


Figure 2.1.10: The *Band search window*.

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

5.47 Enter **5** for *Min. Profiling*, **2** for *Gray zone*, and press <***Preview***>.

The red lines across the lane at the bottom of the window indicate which peaks will be assigned bands with the current settings. The purple area along the curve indicates the minimum profile above which a band is assigned, while the blue area indicates the gray zone in which bands are considered "uncertain".

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

Bands that are found are marked with a green horizontal line, whereas uncertain bands are marked with a small green ellipse (see Figure 2.1.11). If you see any incorrect band assignments you can edit them manually:

- To add a band, hold down the **Ctrl**–key and click on the spot. The cursor automatically jumps to the closest peak; to prevent this, hold down the **Tab**–key while clicking.
- To select a group of bands, hold down the **Shift**–key and click while dragging the mouse pointer diagonally across the bands.
- To delete one or more selected bands, press the **Del**–key.
- To mark a band as uncertain, click on the band and select *Bands* > *Mark band(s) as uncertain* or press **F5**. To mark a band as certain, click on the band and select *Bands* > *Mark band(s) as certain* or press **F6**.

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

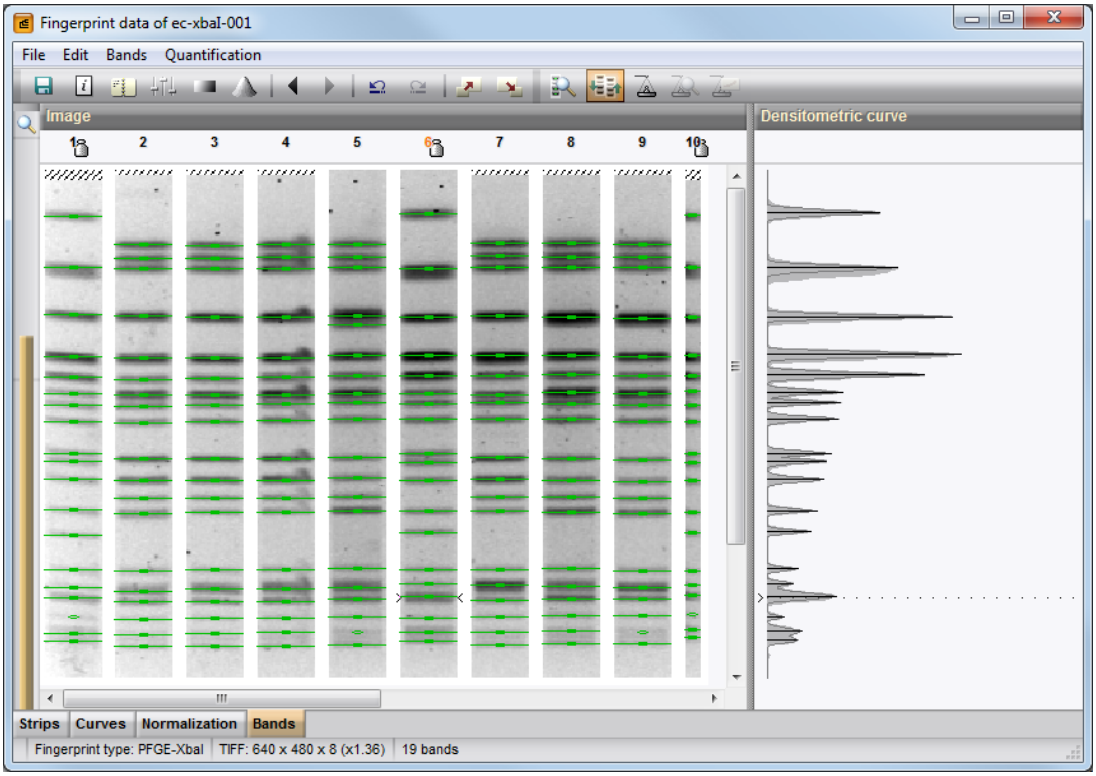


Figure 2.1.11: The *Bands tab*.

5.50 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***>.

5.51 Exit the *Fingerprint data window* by selecting *File* > *Exit*.

5.52 The software asks: "Settings have been changed. Do you want to use the current settings as new defaults?" Select <***Yes***> so that the settings used for this gel will be saved in the fingerprint type settings.

Congratulations! You have processed your first gel. The reference system and fingerprint settings that are saved with the fingerprint type experiment will make future XbaI gels much faster to process.

2.1.6 Link fingerprint data to entries

Although we have created individual fingerprint lanes from our gel image, the software does not know which lanes correspond to which entries in the database. Our next task is to link the fingerprint lanes to the entries in our database.

6.1 In the *Files panel*, double–click on the filename **ec-XbaI-001** to open the *Fingerprint file window* (see Figure 2.1.12).

6.2 Select lane **2** and select *Database* > *Link lane* or press .

6.3 In the dialog box enter **CL004** and press <***OK***>.

You can also link a lane to a database entry by dragging the gray arrow icon to the entry key in the database window. When a lane is linked, the icon becomes purple: .

6.4 Drag the arrow icon of lane **3** to entry **CL001**.

6.5 Continue linking the remaining non–reference lanes from **ec-XbaI-001** to the appropriate database entries as shown in Figure 2.1.12.

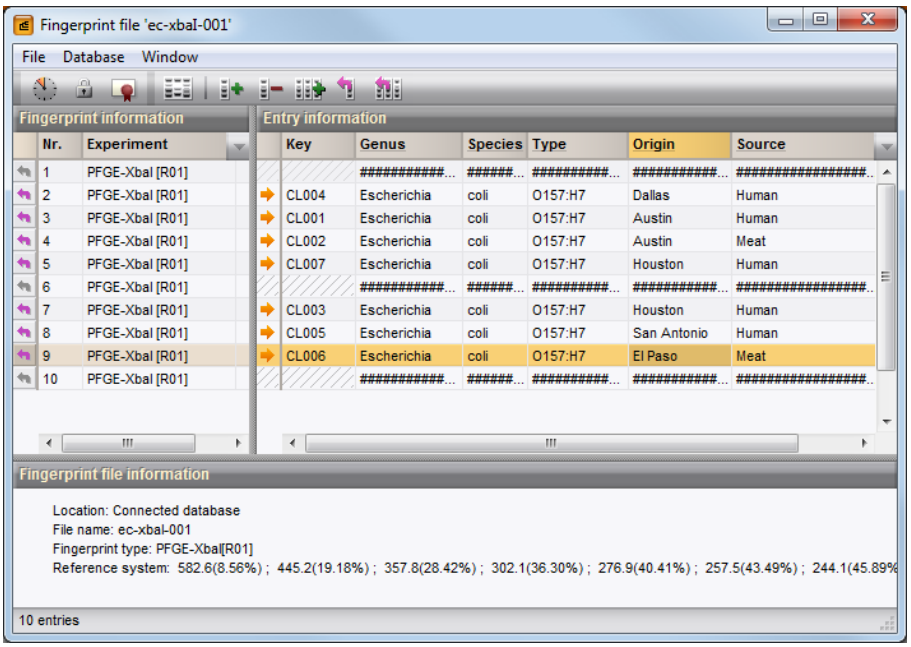


Figure 2.1.12: Fingerprint file with linked lanes.

6.6 After linkage, close the *Fingerprint file window* and open the gel strip for one of the entries in the database by clicking on a colored dot in the *Experiment presence panel*.

The card is displayed in raw mode (i.e. not normalized). The band sizes are shown as relative distances from the top (see Figure 2.1.13).

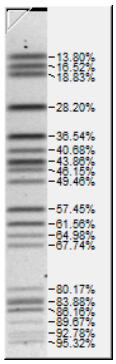


Figure 2.1.13: Fingerprint type experiment card: raw mode.

6.7 Close the card by clicking in the small triangle–shaped button in the upper left corner.

2.1.7 Fingerprint type experiment settings

2.1.7.1 Assigning a standard pattern

7.1 Open the fingerprint experiment type **PFGE-XbaI** by double–clicking on the experiment type in the *Experiments panel*.

The *Fingerprint type window* shows the defined reference positions in relation to the distance on the pattern in percentage. The reference system is called **R01**. The panel left from the reference system is still blank: the fingerprint still misses a standard pattern. We will link a **Standard** pattern (e.g. lane **6** of **ec-XbaI-001**, i.e. the one used for defining the reference system) to the **PFGE-XbaI** fingerprint type as follows:

7.2 Close the *Fingerprint type window*.

7.3 In the *Files panel*, double–click on **ec-XbaI-001** to open the *Fingerprint file window*.

7.4 In the *Fingerprint file window*, add lane **6** to the database by selecting it and pressing .

7.5 In the dialog box, enter ”REF” and press <***OK***>.

7.6 Select <***Yes***> to create the new entry in the database.

A new entry **REF** is created with the pattern from lane **6** of **ec-XbaI-001** linked to it.

7.7 Close the *Fingerprint file window*.

7.8 Open the fingerprint experiment type **PFGE-XbaI** by double–clicking on the experiment type in the *Experiments panel*.

7.9 Press next to **Standard** in the *Settings panel* and drag it over to the **REF** database entry.

The standard pattern is now displayed in the panel next to the reference positions, and the database entry key **REF** is indicated as the **Standard** (see Figure 2.1.14).

If a standard pattern is assigned to a fingerprint type, this standard pattern is shown in the *Normalization tab* of the *Fingerprint data window* (see Figure 2.1.9) to make visual assignment of bands to the reference positions easier. The choice of a standard has no influence on the normalization process since it is only used as visual aid.

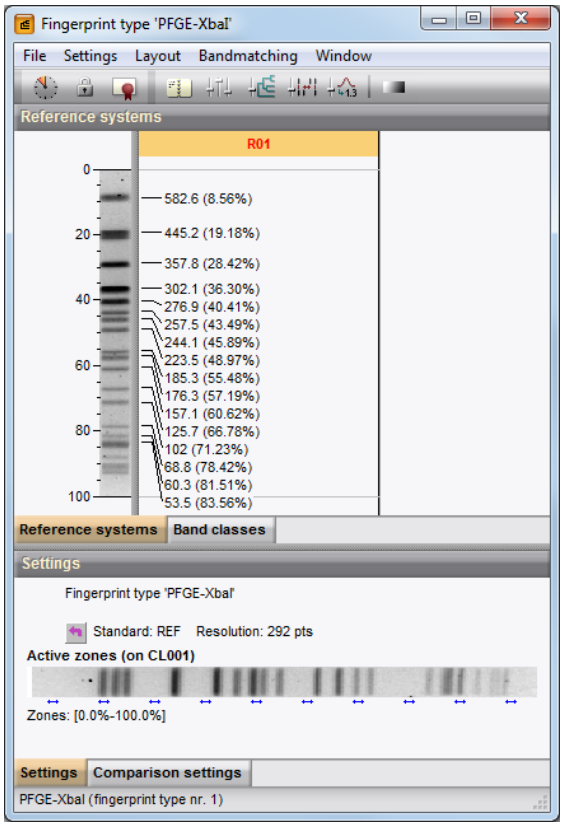


Figure 2.1.14: The *Fingerprint type window*.

2.1.7.2 Calculating a calibration curve

One more step is necessary before we can analyze our fingerprint patterns. Since there are gaps present between the reference system positions, we must tell the software how to convert other band positions into metrics (e.g. molecular weights). We will use the reference system to construct a calibration curve which translates all band positions into metrics.

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

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

7.11 Since the molecular weights were already entered as names for the reference positions, we can copy these molecular weights by selecting *Metrics* > *Copy markers from reference system*. Confirm the action.

7.12 Designate a metric unit with *Metrics* > *Assign units*, enter **kb** and press <***OK***>.

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

The fingerprint type **PFGE-XbaI** is now defined and configured, and one gel has been added to the database.

7.14 Open a gel strip for one of the entries in the database by clicking on a colored dot in the *Experiment presence panel*.

The cards are now displayed in normalized mode (see Figure 2.1.16). Band sizes are shown as molecular sizes based on the regression curve calculated in the previous step.

7.15 Increase or decrease the size of the card using the keyboard by pressing the numerical "+" key (increase) or the numerical "-" key (decrease).

7.16 Close the card by clicking in the small triangle–shaped button in the upper left corner.

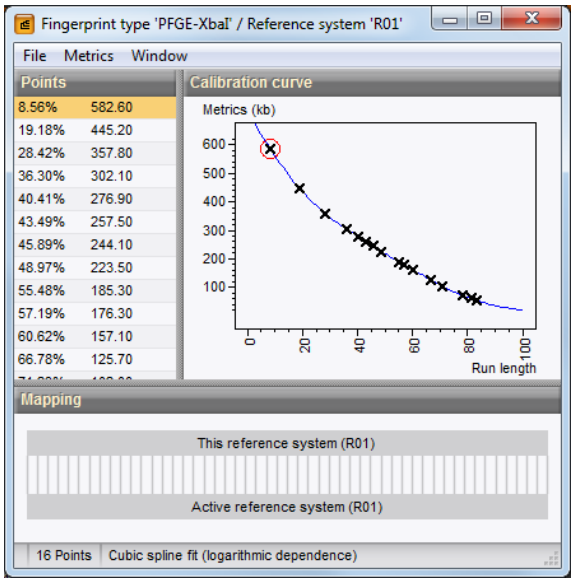


Figure 2.1.15: Calibration curve calculated.

Figure 2.1.16: Fingerprint type experiment card: normalized mode.

2.1.8 Additional practice

2.1.8.1 XbaI-002

A second gel file, `ec-XbaI-002.tif`, which was generated by PFGE with the restriction enzyme **Xba-I** is present in the `BioNumerics Tutorial data \PFGE TIFFS` folder. For this second PFGE-XbaI gel, we will use the same fingerprint type, reference system, and conversion settings used for the first gel.

8.1 Add and process gel `ec-XbaI-002.tif` (see Instruction 4.1 to Instruction 5.52). Use lanes **1**, **4**, **7**, and **10** as the reference lanes. The reference system defined for the first gel is displayed in the *Normalization tab*. In the *Normalization tab*, you will just have to select the reference lanes, and let the software look for the reference bands based on the reference system (Instruction 5.34 to Instruction 5.37 can be skipped).

8.2 After processing the gel, link the lanes to the appropriate database entries (see Figure 2.1.17).

8.3 When linking lane **6** to entry with key **CL010**, BioNumerics will ask whether or not you want to create a duplicate key for this entry.

8.4 Press <***Yes***> to create a duplicate entry for this entry.

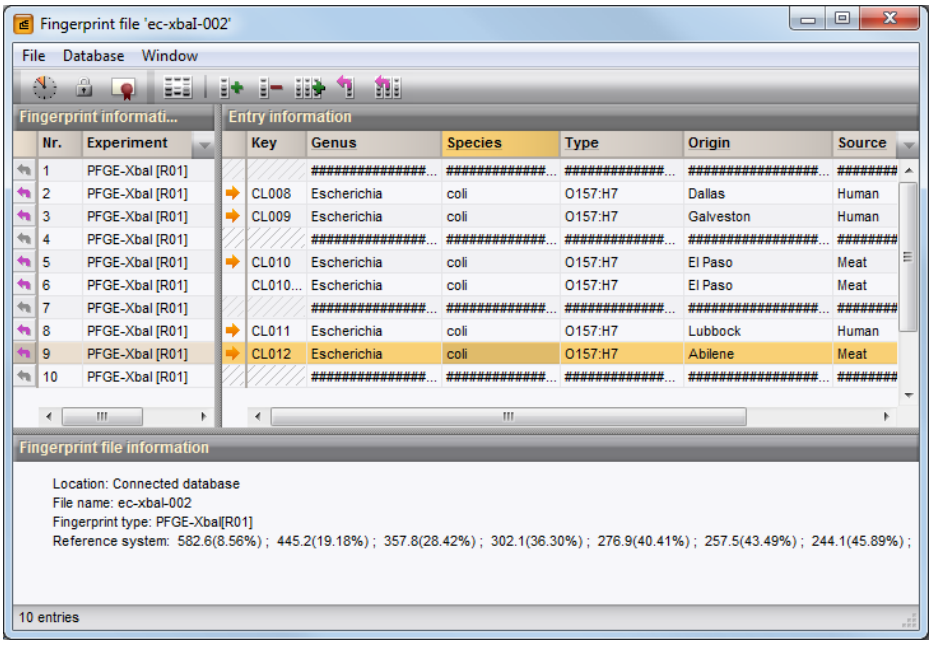


Figure 2.1.17: Linking the lanes for gel ec-XbaI-002 to the database entries.

2.1.8.2 AvrII-001

A second set of gels is present in the `BioNumerics Tutorial data\PFGE TIFFS` folder. These gels are created with the restriction enzyme **Avr-II**. To import these gels in the database, another fingerprint type experiment is needed.

8.5 Create a new fingerprint type experiment called **PFGE-AvrII**, using the same settings as PFGE-XbaI (see Instruction 3.2 to Instruction 3.8).

8.6 Process the AvrII gel `ec-AvrII-001.tif`, using lanes **1**, **6**, and **10** as the reference lanes. Be sure to select **PFGE–AvrII** as the fingerprint type when opening the file for the first time.

The reference system used in the gels XbaI-001 and XbaI-002 is the same as used in gel AvrII-001.

8.7 To facilitate the assignment of reference positions in gel AvrII-001, select *References* > *Copy normalization* in step 3 of one of the XbaI gels and then select *References* > *Paste normalization* in step 3 of the AvrII-001 gel. The reference positions will be transferred from the XbaI gel to the AvrII gel.

8.8 Select *Normalization* > *Auto assign* or press .

8.9 Make sure that *Using bands* is selected and press <***OK***>. Reassign the bands if needed.

After processing the gel, link the lanes to the appropriate database entries as shown in Figure 2.1.18.

8.10 Open the fingerprint type experiment **PFGE-AvrII** by double–clicking on the experiment type in the *Experiments panel*.

8.11 Press next to **Standard** in the *Settings panel* and drag it over to the **REF** database entry.

8.12 In the same window, select *Settings* > *Edit reference system* and select *Metrics* > *Copy markers from reference system*. Confirm the action.

8.13 Designate a metric unit with *Metrics* > *Assign units*, and enter **kb**. Press <***OK***>.

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

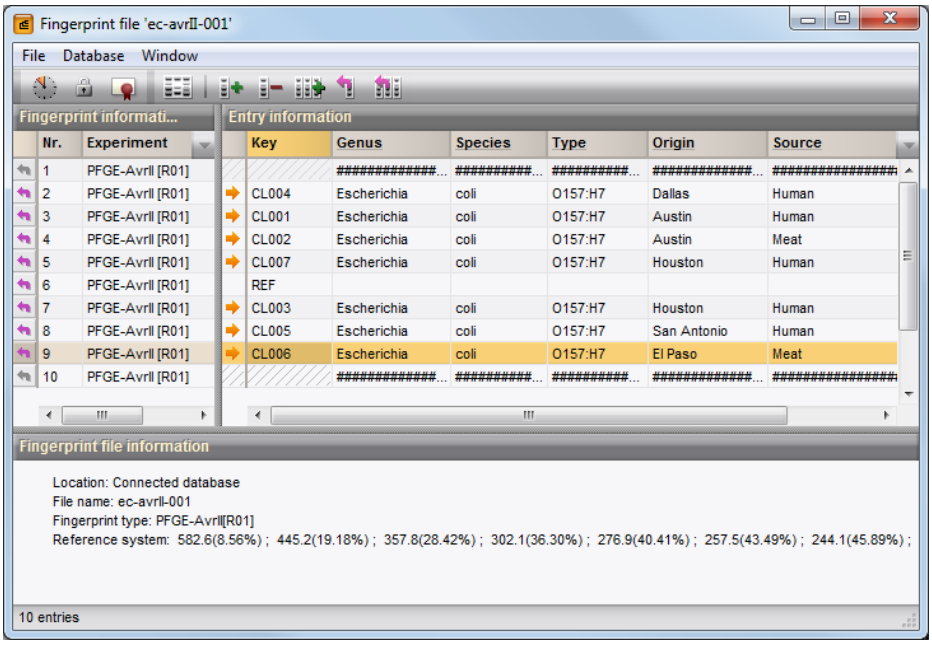


Figure 2.1.18: Linking the lanes for gel ec-AvrII-001 to the database entries.

2.1.8.3 AvrII-002

Process the second PFGE-AvrII gel, `ec-AvrII-002.tif`, using lanes **1**, **4**, **7**, and **10** as the reference lanes. Link the lanes to the database entries as you did for XbaI-002 (see Figure 2.1.17).

Chapter 2.2

Character data

2.2.1 Introduction

In this Chapter we will:

- Create a character type experiment
- Import an Excel character data file
- Import a text character data file
- Optimize the character type experiment settings

2.2.2 Sample data

As an exercise, we will import data from a text and Excel file in our **E. coli** database. If the *Install sample and tutorial data* feature was checked in the *Install wizard* (see Figure 1.1.4), these files can be found in the `Sample and Tutorial data` folder in the database home directory. Alternatively, these files can be downloaded from our website: go to `http://www.applied-maths.com/download/sampledata.htm` and click on "BioNumerics Tutorial Data".

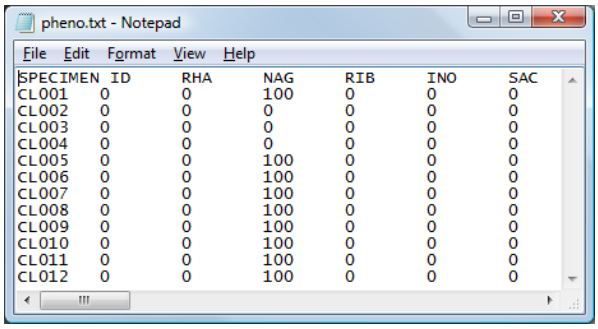
The Excel file `Ecoli-Biolog.XLS`, located in the `BioNumerics Tutorial data\Char import` folder, contains data from 96-well micro titer plates measuring the metabolic activity for various carbon sources (Biolog system). The file contains twelve samples, CL001 to CL012, corresponding to the twelve *E. coli* strains already present in the database (see Figure 2.2.1).

	A	B	C	D	E	F	G
1	Strain Number	Blank	a-Cyclodextrin	Dextrin	Glycogen	Tween 40	Tween 80
2	CL001	9	8	12	13	17	15
3	CL005	7	2	15	9	8	13
4	CL011	3	10	20	19	18	23
5	CL012	6	9	25	15	8	17
6	CL008	5	8	19	11	21	19
7	CL007	11	4	13	4	11	14
8	CL002	10	6	15	12	14	21
9	CL006	7	4	20	5	9	13
10	CL004	8	10	14	7	20	16
11	CL010	10	5	27	21	15	23
12	CL003	9	0	13	3	12	16
13	CL009	10	3	18	15	5	14
14							

Ecoli-biolog

Figure 2.2.1: Part of the Excel file.

The text file `pheno.txt`, located in the `BioNumerics Tutorial data \Char import` folder, contains biochemical data. The samples, CL001 to CL012, correspond to the twelve *E. coli* strains already present in the database (see Figure 2.2.2).



pheno.txt - Notepad

File Edit Format View Help

SPECIMEN	ID	RHA	NAG	RIB	INO	SAC
CL001	0	0	100	0	0	0
CL002	0	0	0	0	0	0
CL003	0	0	0	0	0	0
CL004	0	0	0	0	0	0
CL005	0	0	100	0	0	0
CL006	0	0	100	0	0	0
CL007	0	0	100	0	0	0
CL008	0	0	100	0	0	0
CL009	0	0	100	0	0	0
CL010	0	0	100	0	0	0
CL011	0	0	100	0	0	0
CL012	0	0	100	0	0	0

Figure 2.2.2: Part of the text file.

2.2.3 Creating a character type experiment

A new character type experiment needs to be created in our **E. coli** database before we can import character data from the Excel sample file (see Figure 2.2.1) into the database.

3.1 In the BioNumerics startup screen, double-click on the **E. coli** database - created in 1.2.1 - to open it.

3.2 In the *BioNumerics main window*, select *Experiments* > *Create new character type* from the main menu, or press the button from the *Experiments panel toolbar* and select *New character type*.

3.3 Enter a name, for example **Biolog** and press <***Next***>.

3.4 Since the tests in our example Excel file (see Figure 2.2.1) differ in intensity, select *Numerical values* for the kind of character data. Since the values are integers, enter 0 for the number of decimals to use. Press <***Next***>.

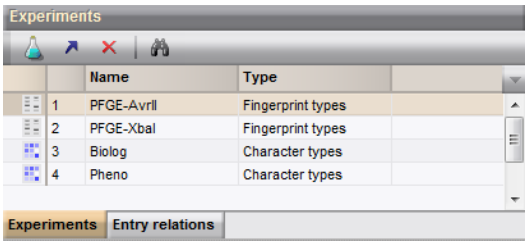
3.5 The **Biolog** character type has a closed (fixed) set of characters, so select <***No***> in the next step and leave the *Layout* as it is.

3.6 Press <***Finish***> to complete the setup of the new character type.

A second character type experiment needs to be created in our **E. coli** database for the import and storage of the data that is present in the sample text file (see Figure 2.2.2).

3.7 Create a new character type experiment called **Pheno**, using the same settings as the **Biolog** experiment.

Two character types are now listed in the *Experiments panel* (see Figure 2.2.3).



Experiments

	Name	Type
1	PFGE-AvrII	Fingerprint types
2	PFGE-XbaI	Fingerprint types
3	Biolog	Character types
4	Pheno	Character types

Experiments | Entry relations

Figure 2.2.3: The *Experiments panel*.

2.2.4 Importing character data from external files

2.2.4.1 Importing character data from an Excel file

4.1 Select *File* > *Import* or press to call the Import tree.

4.2 Select *Character type data* in the import tree, highlight *Import fields and characters (ODBC)* and press <***Import***>.

4.3 Press the <***Build***> button in the *ODBC connection* step.

The dialog box that pops up is generated by your Windows operating system and may differ depending on the Windows version installed.

4.4 Click on the tab "Machine Data Source", pick "Excel Files" from the list and press <***OK***>. If the data source is not listed, create a new data source.

4.5 Browse for the `Ecoli-Biolog.XLS` file in the `BioNumerics Tutorial data \Char import` folder, select the file and press <***OK***>.

The ODBC string is updated in the *ODBC connection string* input box.

4.6 Select *Ecoli-Biolog$* from the *Table* list and press <***Next***> (see Figure 2.2.4).

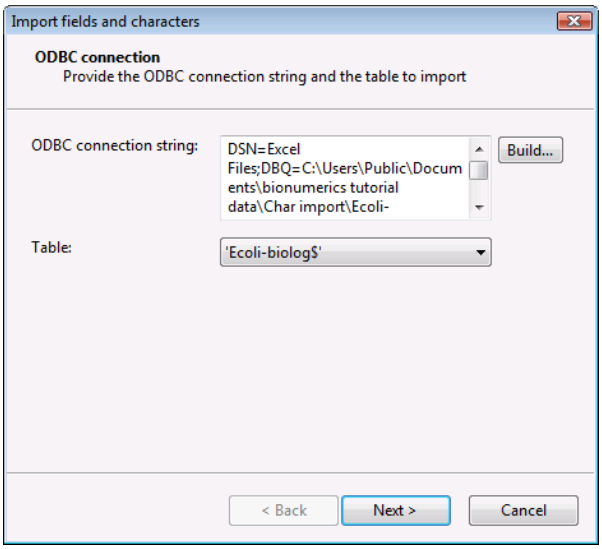


Figure 2.2.4: Select the file to import and the table.

4.7 Press the <***Create new***> button to create a new import template.

This brings up a new dialog box (see Figure 2.2.5). Each column in the selected file corresponds to a row in the grid (column 1 in the file corresponds to row 1 in the grid, column 2 corresponds to row 2, etc.). The text *File field* is specified in the *Source type* column and the column names are displayed in the *Source* column.

4.8 Select the first row entry in the grid, press the <***Edit destination***> button and select the BioNumerics *Key* field from the list. Press <***OK***>.

The grid is updated (see Figure 2.2.5).

4.9 Select the second row entry, hold the Shift-key, scroll down the list and select the last row in the grid. All rows holding character information should now be selected in the grid panel.

4.10 Press the <***Edit destination***> button and select the *Biolog* option that is listed under the topic *Character* (see Figure 2.2.6). Press <***OK***>.

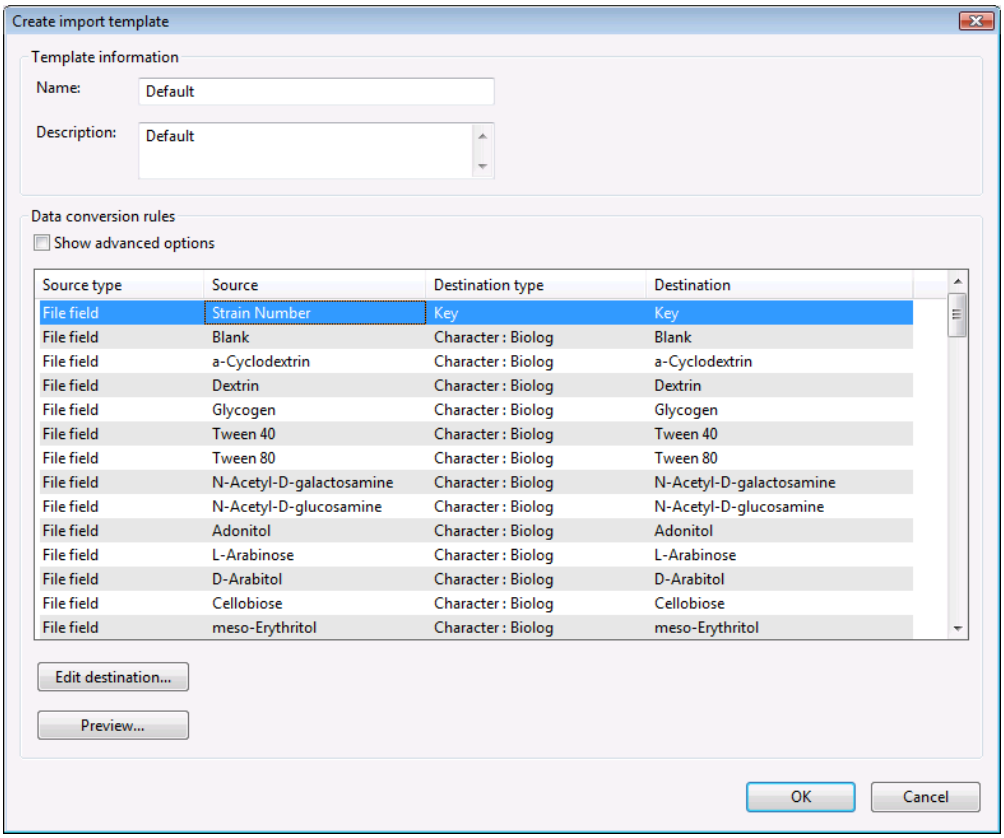


Figure 2.2.5: Define a new import template.

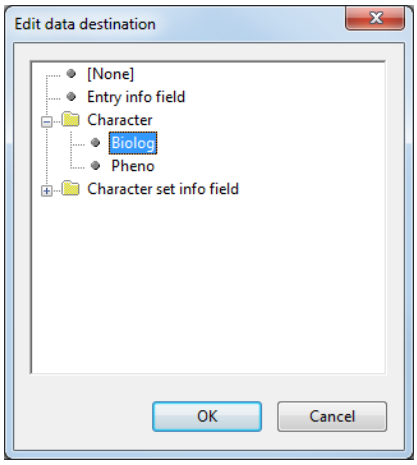


Figure 2.2.6: Link character information to the Biolog experiment.

4.11 Press <***OK***> once more to accept the default suggested names and press <***Yes***>.

The grid is updated (see Figure 2.2.5).

4.12 Optionally, change the default suggested template *Name* and press <***OK***>.

The import template is added to the list and is automatically selected.

4.13 Press <***Next***> twice.

After import, the new characters have been added to the **Biolog** character type and the character data is

linked to entries CL001 to CL012.

4.14 Open the **Biolog** experiment to verify that new characters have been imported (double-click on the experiment in the *Experiments panel*).

96 characters are listed in the **Biolog** *Character type window* (see Figure 2.2.7).

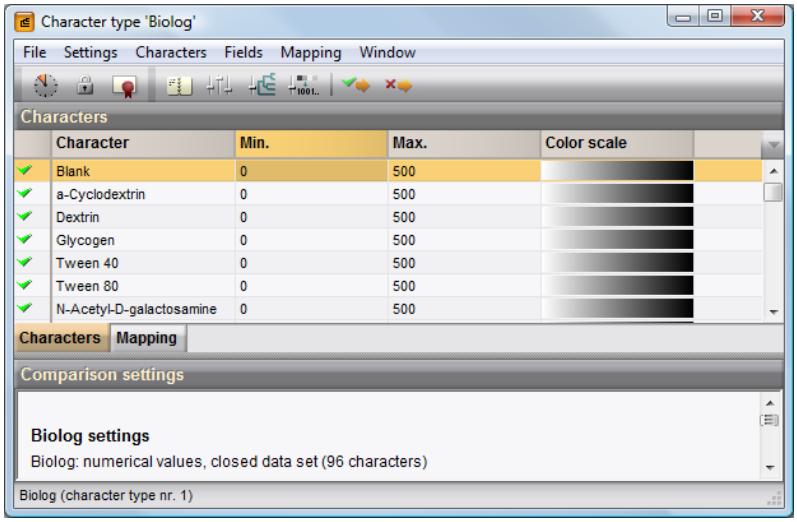


Figure 2.2.7: Characters listed in the *Character type window*.

4.15 Close the *Character type window*.

2.2.4.2 Importing character data from a text file

4.16 Select *File* > *Import* or press to call the Import tree.

4.17 Select *Character type data* in the Import tree, highlight *Import fields and characters (text file)* and press <***Import***>.

4.18 Press <***Browse***> and navigate to the path where the `pheno.txt` file is stored (`BioNumerics Tutorial data \Char import`). Select the file.

4.19 Select *TAB* as the *Field separator* and press <***Next***> (see Figure 2.2.8).

4.20 Press the <***Create new***> button to create a new import template.

This brings up a new dialog box (see Figure 2.2.9). Each column in the selected file corresponds to a row in the grid (column 1 in the file corresponds to row 1 in the grid, column 2 corresponds to row 2, etc.). The text *File field* is specified in the *Source type* column and the column names are displayed in the *Source* column. The last row in the grid holds the file name.

4.21 Select the first row entry in the grid, press the <***Edit destination***> button and select the BioNumerics *Key* field from the list. Press <***OK***>.

The grid is updated (see Figure 2.2.9).

4.22 Select the second row entry, hold the Shift-key, scroll down the list and select the last row entry in the grid that has the text *File field* displayed in the first column (make sure the last row in the grid is not selected). All rows holding character information should now be selected in the grid.

4.23 Press the <***Edit destination***> button and select the *Pheno* option that is listed under the topic *Character*. Press <***OK***>.

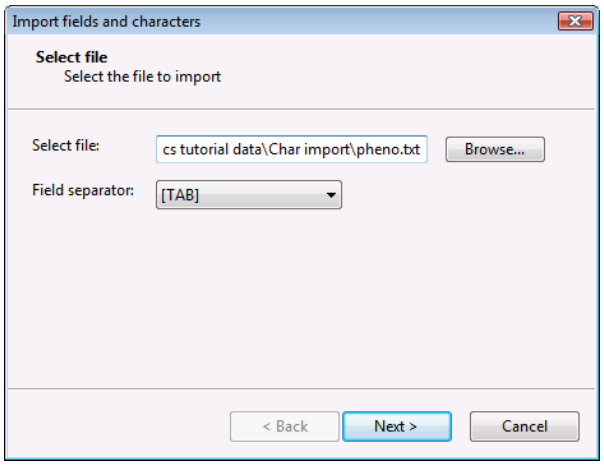


Figure 2.2.8: Select the file to import and the separator.

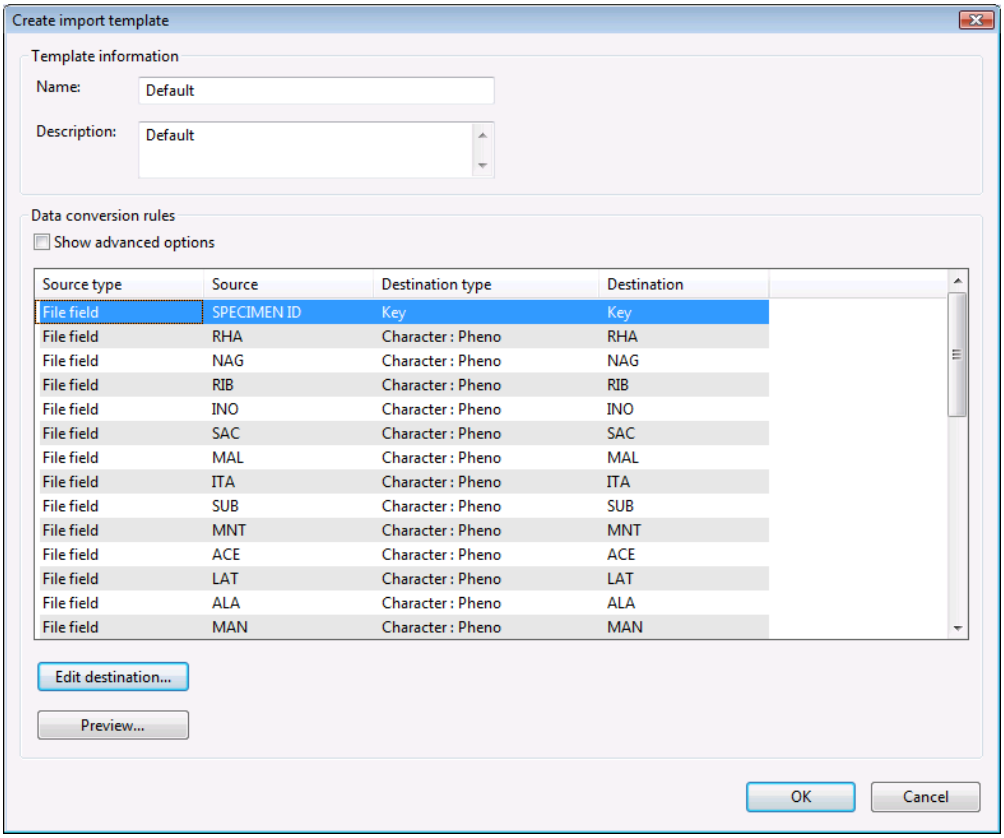


Figure 2.2.9: Define a new import template.

4.24 Press <***OK***> once more to accept the default suggested names and press <***Yes***>.

The grid is updated (see Figure 2.2.9).

4.25 Optionally, change the default suggested template *Name* and press <***OK***>.

The import template is added to the list and is automatically selected.

4.26 Press <***Next***> twice.

After import, the new characters have been added to the **Pheno** character type and the character data is

linked to entries CL001 to CL012.

4.27 Open the **Pheno** experiment to verify that 34 characters have been imported (double-click on the experiment in the *Experiments panel*).

4.28 Close the *Character type window*.

2.2.5 Changing the character type settings

The settings that were chosen when the character type was created can be changed at any time in the *Character type window*. The way the characters are displayed can also be tailored in this window.

2.2.5.1 Experiment card settings

Character data can be displayed as a list of numerical values or as a graphical representation of the original experiment.

5.1 In the *Database entries window*, click on a colored dot in the *Experiment presence panel* representing the **Biolog** experiment for any database entry.

The experiment card opens. The layout of the card can be changed in the *Character type window*.

5.2 Press the top left corner of the experiment card to close it.

5.3 Double-click on the **Biolog** experiment in the *Experiments panel*.

5.4 Select *Settings* > *General settings* in the *Character type window*.

5.5 Press the *Experiment card tab* to define some graphical attributes of the experiment card.

5.6 Select *Plate* to represent the experiment card as a micro titer plate, enter **12** as the number of columns and select *Large circular cup* as the cell type.

5.7 After having specified these new settings (see Figure 2.2.10), press <***OK***>. Close the *Character type window*.

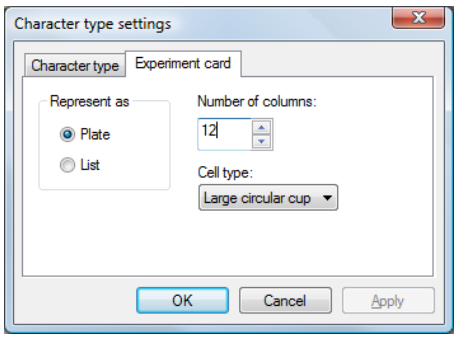


Figure 2.2.10: The *Character type settings dialog box*.

5.8 In the *Database entries window*, click on a colored dot in the *Experiment presence panel* representing the **Biolog** experiment for any database entry.

The experiment card now shows the characters displayed as a 12x8 plate (see Figure 2.2.12).

5.9 Press the triangle in the top left corner of the experiment card to close it.

2.2.5.2 Character color setup

Character data can be represented as colors along a spectrum of infinitely adjustable colors. This flexibility in presentation is especially helpful when viewing character data in the *Comparison window* (see 3.3).

5.10 Double-click on the **Biolog** experiment in the *Experiments panel*.

5.11 Select a character from the list, and select *Characters* > *Change character color scale*.

5.12 Click on the left side of the color scale (negative reaction).

The left side of the color scale should now be marked with a black triangle.

5.13 Adjust the red, green and blue components by sliding the switches until you have obtained the desired color for a negative reaction (e.g. red).

5.14 Click on the right side of the color scale (positive reaction) and adjust the red, green and blue components until you have obtained the desired color (e.g. green).

5.15 To add more transition colors, press <***Add point***>.

5.16 Adjust the components until you have obtained the desired color (e.g. yellow). Press <***OK***>.

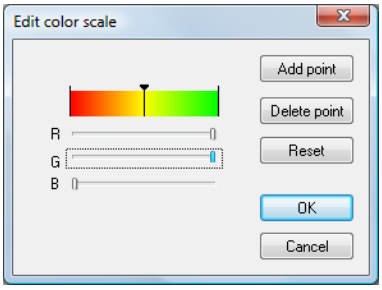


Figure 2.2.11: The *Edit color scale dialog box*.

5.17 Select *Characters* > *Copy color scale to all other characters* to copy the defined color scale to all characters. Confirm the action.

All **Biolog** characters now have the same color scale.

5.18 Close the *Character type window*.

5.19 In the *Database entries window*, click on a colored dot in the *Experiment presence panel* representing the **Biolog** experiment for any database entry.

The experiment card shows the characters using the defined color scale (see Figure 2.2.12).

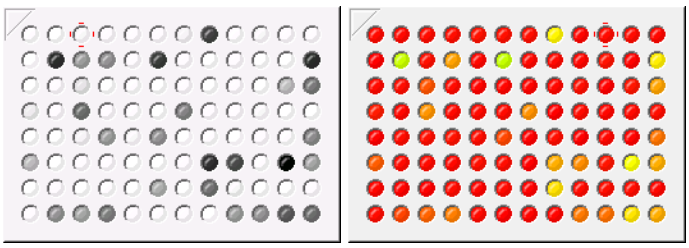


Figure 2.2.12: Left: default color scale, Right: user-defined color scale.

Chapter 2.3

Sequence data

2.3.1 Introduction

In this Chapter we will:

- Create a sequence type experiment
- Import sequences from a FASTA file
- Import chromatogram trace files

2.3.2 Creating a sequence type experiment

2.1 In the BioNumerics startup screen, double-click on the **E. coli** database - created in 1.2.1 - to open it.

2.2 In the *BioNumerics main window*, select *Experiments* > *Create new sequence type* from the main menu, or press the button from the *Experiments panel toolbar* and select *New sequence type*.

2.3 Enter **Ribosomal** in the wizard and press <***Next***>.

2.4 Select *Nucleic acid sequences* and press <***Finish***> to complete the setup of the new sequence type.

The **Ribosomal** sequence type is now listed in the *Experiments panel* (see Figure 2.3.1).

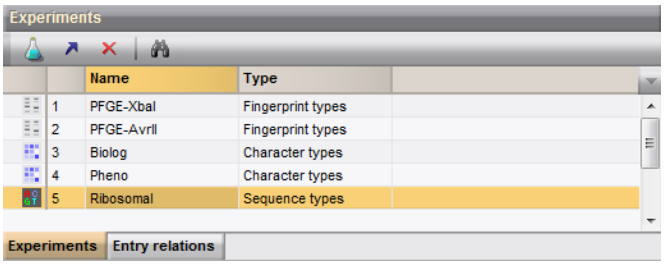


Figure 2.3.1: The *Experiments panel*.

2.3.3 Importing FASTA sequences

Sequence data stored in GenBank, EMBL or FASTA format in a text file can be imported in the database using the BioNumerics import functions. As an exercise, we will import the sequences that are present in the FASTA-formatted text file `Ecoli-seq.txt` into our **E. coli** database. If the *Install sample and tutorial data* feature was checked in the *Install wizard* (see Figure 1.1.4), this text file can be found in the `Sample and Tutorial data` folder in the database home directory. Alternatively, the file can be downloaded from our website: go to `http://www.applied-maths.com/download/sampledata.htm` and click on "BioNumerics Tutorial Data". The file contains sequences for twelve samples, CL001 to CL012, corresponding to the twelve *E. coli* strains already present in the database.

3.1 Select *File* > *Import* or press to call the Import tree.

3.2 Select *Sequence type data* in the import tree, highlight *Import sequences from text files* and press <***Import***>.

3.3 Press the <***Browse***> button.

3.4 Browse for the `Ecoli-seq.txt file` in the `BioNumerics Tutorial data \Sequences \FASTA` folder, and select the file. Press <***Next***>.

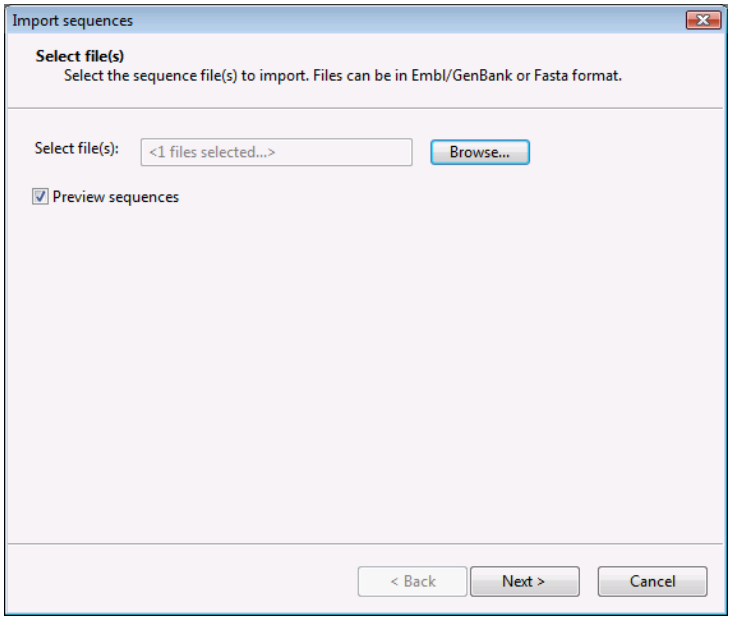


Figure 2.3.2: Select the sequence file to import.

If the option <***Preview sequences***> was checked in the first step of the wizard, the second step displays all sequences found in the selected file (see Figure 2.3.3). The *File name* column holds the name of the selected file, the *Length* column displays the size of the sequences, and the *Header* column holds the information that is present in the description line.

3.5 Press <***Next***>.

3.6 Press the <***Create new***> button to create a new import template.

This brings up a new dialog box (see Figure 2.3.4). When sequences are stored in FASTA format, each sequence begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater than (">") symbol. The description line contains the *FASTA tags*, separated by a pipe ("|") symbol. In the example text file only one FASTA tag is present in the description line of each sequence. This tag corresponds to the *Key* information in the database.

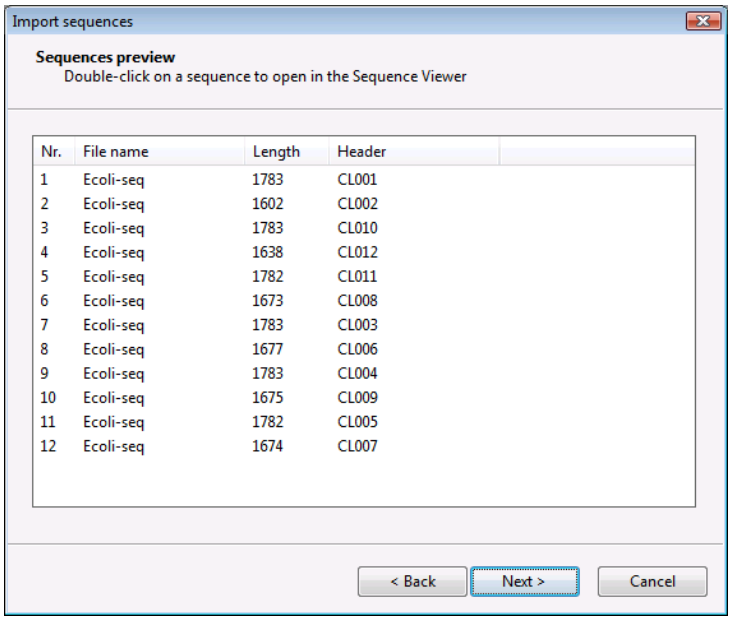


Figure 2.3.3: Preview.

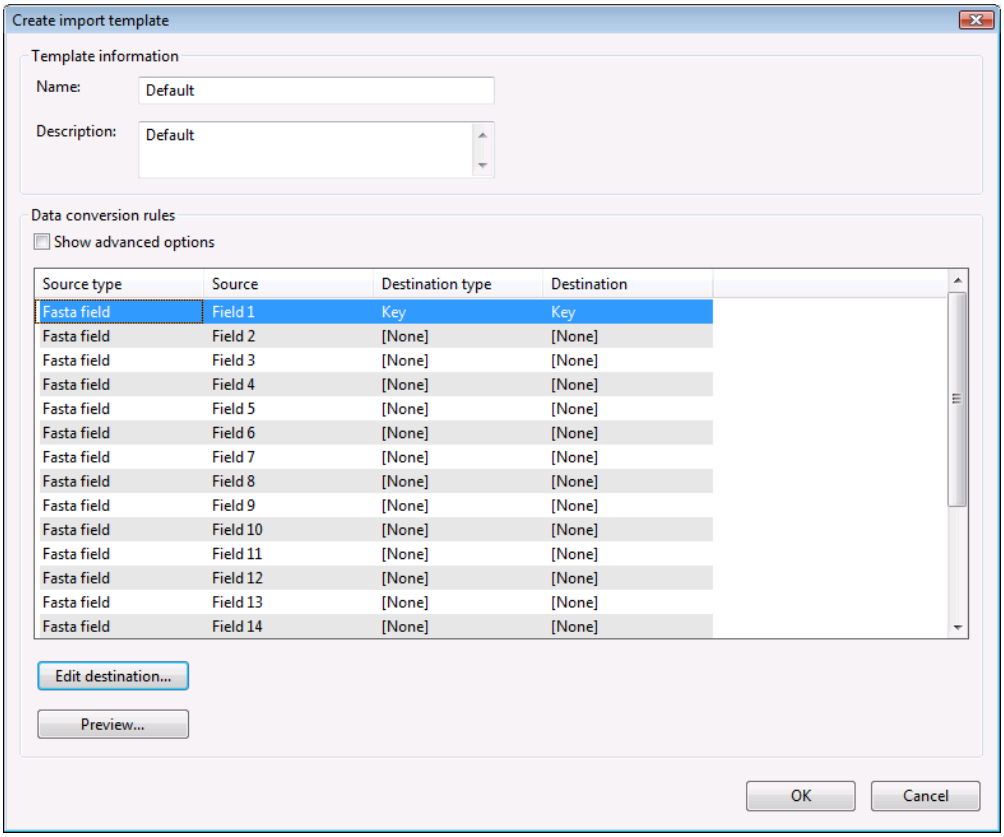


Figure 2.3.4: Define a new import template.

3.7 Select the first row entry in the grid, press the <***Edit destination***> button and select the BioNumerics *Key* field from the list. Press <***OK***>.

The grid is updated (see Figure 2.3.4).

3.8 Press <***OK***>.

The import template is added to the list and is automatically selected.

3.9 Press <***Next***>.

3.10 In the next step, select the **Ribosomal** sequence type from the *Sequence experiment* list. Leave all other settings unaltered and press <***Next***>.

The sequences are imported in the database and are linked to the *E. coli* entries in the database.

3.11 In the *Database entries window*, click on a colored dot in the *Experiment presence panel* representing the **Ribosomal** experiment for any database entry while holding the Shift-key.

The experiment card shows the imported sequence for the selected database entry (see Figure 2.3.5).

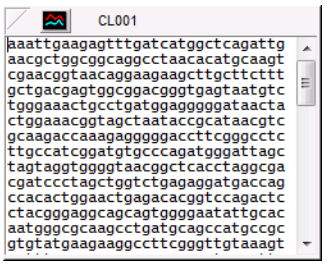


Figure 2.3.5: The *Sequence experiment card* for entry with key CL001.

3.12 Press the triangle in the top left corner of the experiment card to close the card.

3.13 In the *Database entries window*, click on a colored dot in the *Experiment presence panel* representing the **Ribosomal** experiment for any database entry this time without holding the Shift-key.

The *Sequence Viewer* pops up. Frame analysis, restriction enzyme analysis, and primer analysis can be executed from this window.

3.14 Close the *Sequence Viewer*.

2.3.4 Importing chromatogram trace files

Assembler is a BioNumerics program that assembles contig sequences from partial sequences. The program accepts flat text files as well as binary chromatogram files (Amersham, Applied Biosystems, Beckman).

As an example, we will import trace files from the influenza A virus strains into a new database. This example dataset can be found on our website: go to `http://www.applied-maths.com/download/sampledata.htm` and click on "Batch assembly & Alignment data". If the *Install sample and tutorial data* feature was checked in the *Install wizard* (see Figure 1.1.4), the trace files can also be found in the `Sample and Tutorial data\Example_traces` folder in the database home directory.

4.1 In the BioNumerics Startup screen, create a new database (see 1.2.1). Call it e.g. **SeqAssembly** and leave all settings to their defaults.

In the new, empty created database, install the *Batch sequence assembly plugin*:

4.2 Press *File* > *Install/Remove plugins*.

4.3 Select the *Batch sequence assembly plugin* from the list of plugins, and press <***Install***>.

4.4 Confirm the installation of the plugin and close the *Plugin installation toolbox*.

The trace files from the influenza A virus strains will be imported and assembled in batch using this *Batch sequence assembly plugin*. The steps will not be explained into detail. For detailed information on the use of this plugin, press the <***Manual***> button in the *Plugin installation toolbox*.

4.5 In the *BioNumerics main window*, select *File* > *Batch sequence assembly* > *Batch sequence assembly*.

4.6 Browse for the `Example_traces` folder, and select all .SCF trace files that start with the name **NA** (= neuraminidase). Press <***Open***>.

4.7 In the *File parsing settings dialog box* that appears (see Figure 2.3.6), select *SCF* as *File format*, check *Fetch experiment from file parsing* and press <***Proceed***>.

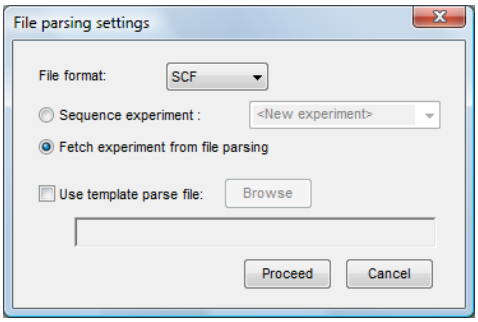


Figure 2.3.6: The *File parsing settings dialog box* from the *Batch sequence assembly plugin*.

4.8 In the *File parsing dialog box*, use "[EXP]_[KEY]–*" as parsing string and press <***Parse***>.

The key and experiment name is parsed from the filename as shown in Figure 2.3.7.

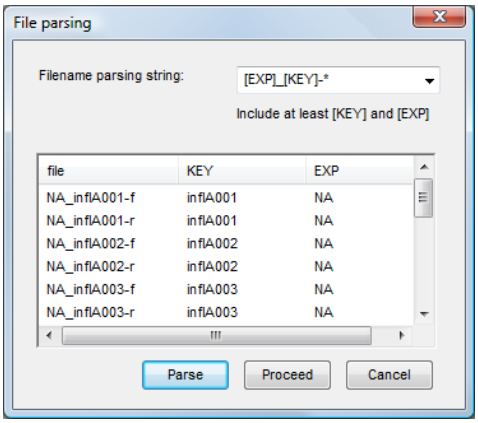


Figure 2.3.7: The *File parsing dialog box* from the *Batch sequence assembly plugin*, showing the key and the experiment name parsed from the filenames in the example data set.

4.9 Press <***Proceed***> to continue.

In the *Experiment settings dialog box* that appears, **NA** is listed under *Experiments missing in database*.

4.10 Press <***Create***> to have the sequence type experiment automatically created by the software. Confirm the action.

4.11 Press <***Trimming settings***> to pop up the *Trimming settings dialog box*.

4.12 For the sequences in the example dataset, enter the trimming settings as specified in Figure 2.3.8.

4.13 Press <***OK***> to close the *Trimming settings dialog box*.

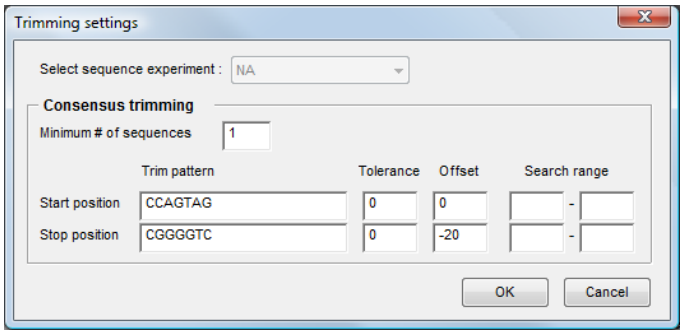


Figure 2.3.8: The *Trimming settings dialog box*, displaying the trimming settings for the NA sequence example data.

4.14 Press <***Proceed***> and then <***Assemble***> to have the sequences automatically assembled by the *Batch sequence assembly plugin*.

4.15 When the assemblies are processed, an interactive report appears (see Figure 2.3.9).

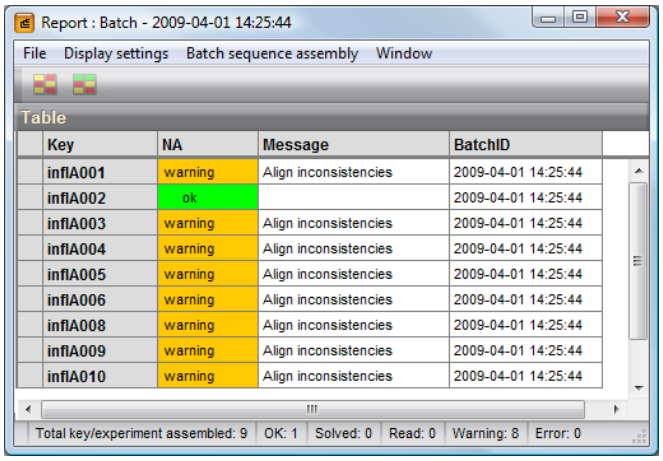


Figure 2.3.9: The *Report window* for the assembled batch of example data.

Warning messages are displayed if alignment inconsistencies occurred that were resolved under the consensus settings.

4.16 Double-click on the *warning* cell of key **inflA001** to display the *Detailed report window* and to launch *Assembler*.

4.17 Select the *Aligned traces tab* to display the two trace chromatograms (see Figure 2.3.10).

4.18 Use the zoom sliders to obtain the best view of the chromatograms and the sequences.

In the Assembly of key **inflA001** there is an inconsistency in the alignment around position 365: a T is missing in the reverse sequence. Using the default Assembler consensus settings, the combination of a gap in one sequence and a nucleotide in the other sequence, will insert a base in the consensus sequence.

4.19 Close the *Assembler window*.

4.20 Double-click on the STATUS cell of the message reporting the alignment inconsistency in the *Detailed report window* (see Figure 2.3.11).

4.21 Check the *Solved* radio button in the next window and press <***OK***>.

4.22 Close the *Detailed report window*.

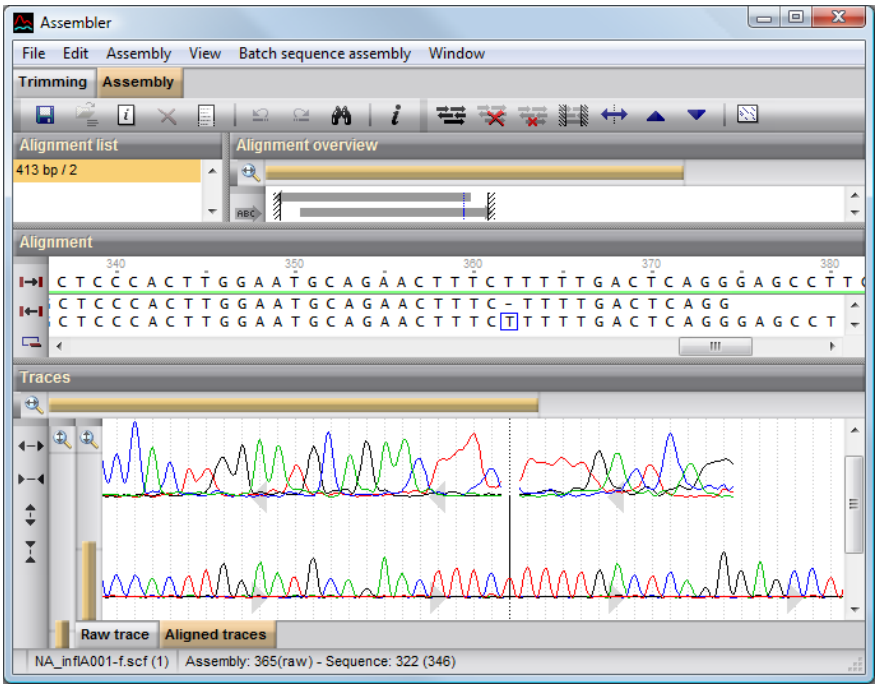


Figure 2.3.10: The *Assembler window*.

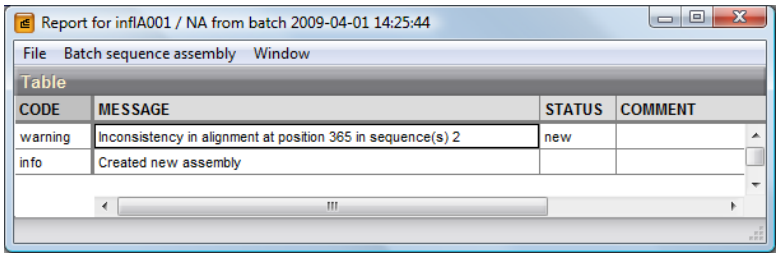


Figure 2.3.11: The *Detailed report window*.

The cell in the **NA** column in the *Report window* will now show *solved* for key **inflA001**.

4.23 Close the *Report window*.

The **NA** sequence type is listed in the *Experiments panel* in the *BioNumerics main window*.

4.24 Click on a colored dot in the *Experiment presence panel* representing the **NA** experiment for database entry **inflA001** while holding the Shift-key.

The experiment card shows the consensus sequence for the selected database entry.

4.25 Select in the experiment card to launch *Assembler* again.

4.26 Close the *Assembler window*.

4.27 Click on a colored dot in the *Experiment presence panel* representing the **NA** experiment for database entry **inflA001** this time without holding the Shift-key.

The *Sequence Viewer* pops up, showing the consensus sequence for the selected database entry.

4.28 To launch Assembler from the *Sequence Viewer*, press the button or use the menu option *File* > *Import sequence using assembler*.

4.29 Close the *Assembler window* and the *Sequence Viewer*.

Part 3

Comparisons

Chapter 3.1

General comparison functions

3.1.1 Comparison settings

If the *Install sample database* feature was checked in the *Install wizard* (see Figure 1.1.4), the database **DemoBase Connected** should be listed in the BioNumerics Startup screen.

1.1 In the BioNumerics Startup screen, double-click on **DemoBase Connected** to open it.

1.2 To view the comparison settings for an experiment in the database (e.g. **RFLP2**) double–click on the experiment in the *Experiments panel*.

The comparison settings are shown in the *Comparison settings panel*.

1.3 Select *Settings* > *Comparison settings* to open the *Comparison settings window*.

1.4 Change the settings if desired and close the *Comparison settings window*.

1.5 Close the *Experiment type window*.

3.1.2 Comparing two entries

A pairwise comparison is useful because it shows every detail of the comparison for each experiment. For example, if you want to know exactly how the bands from two fingerprints are aligned, a pairwise comparison will show the alignment.

2.1 In the BioNumerics startup screen, double-click on **DemoBase Connected** to open it. In case the **DemoBase Connected** was already open, clear any previous selection by pressing **F4**.

2.2 Select any two entries you want to compare (except **STANDARD**). Use the Ctrl-button to select the entries.

2.3 Select *Comparison* > *Compare two entries* or press **Ctrl+Alt+C**.

In the *Pairwise comparison window*, all experiments present in the database are listed in the *Experiments panel*. The similarity values are shown in the *Similarity* column (see Figure 3.1.1).

2.4 Select an experiment in the left panel to display the comparison details in the right panel.

2.5 Close the window.

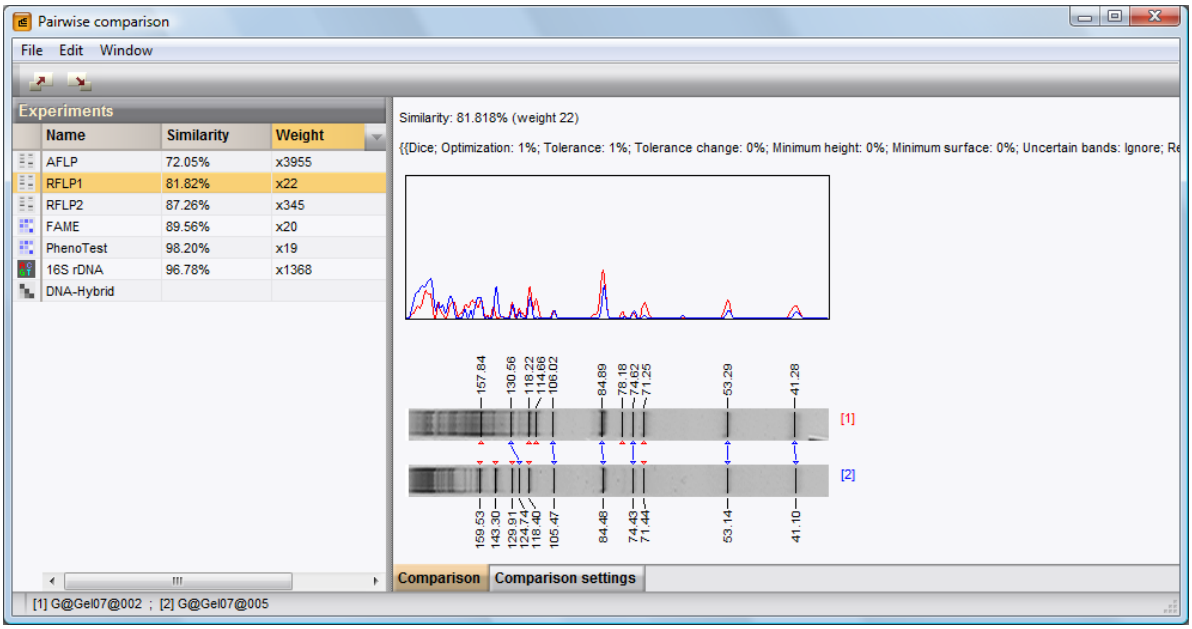


Figure 3.1.1: The *Pairwise comparison window*.

3.1.3 Creating a new comparison

To compare more than two entries, the *Comparison window* is used.

3.1 Clear any previous selection by pressing **F4**.

3.2 Select *Edit* > *Search entries...* (, **F3**).

3.3 Specify the name **STANDARD** in the *Genus* text box, select *Negative search* and press the <***Search***> button (see Figure 3.1.2).

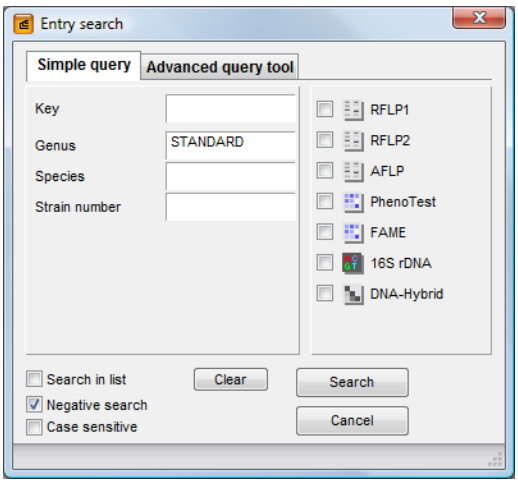


Figure 3.1.2: The *Entry search window*.

All non–STANDARD entries are selected in the database.

3.4 Select *Comparison* > *Create new comparison* (**Alt+C**) or press the button from the *Comparisons panel* toolbar.

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

3.1.4 Comparison window

4.1 Drag the separator lines between the panels in the *Comparison window* to the left or right to size them optimally.

4.2 Drag the separator lines between the field names in the *Information fields panel* to the left or right to size the columns optimally.

The *Comparison window* shows the database entries in the *Information fields panel* and the images of the experiments in the *Experiment data panel*. The *Experiments panel* list the experiment types, the *Groups panel* lists the group sizes and names, and the *Analyses panel* displays the analyses. The other two panels contain the dendrogram (*Dendrogram panel*) and similarity matrix (*Similarities panel*), if calculated.

3.1.4.1 Comparison layout

Each experiment type in the *Experiments panel* contains two objects: a button and the experiment type name on the right hand side of the button. In case of a fingerprint type, the button is shown as ; character experiments as ; sequence types as ; matrix types as ; trend data types as and composite data sets as .

4.3 Press next to **RFLP1**. The fingerprints for **RFLP1** are shown in the *Experiment data panel* (see Figure 3.1.3).

When the experiments are shown, the icon is displayed with a green check.

4.4 Press next to **RFLP2**. The **RFLP2** data are shown to the right of **RFLP1** (see Figure 3.1.3).

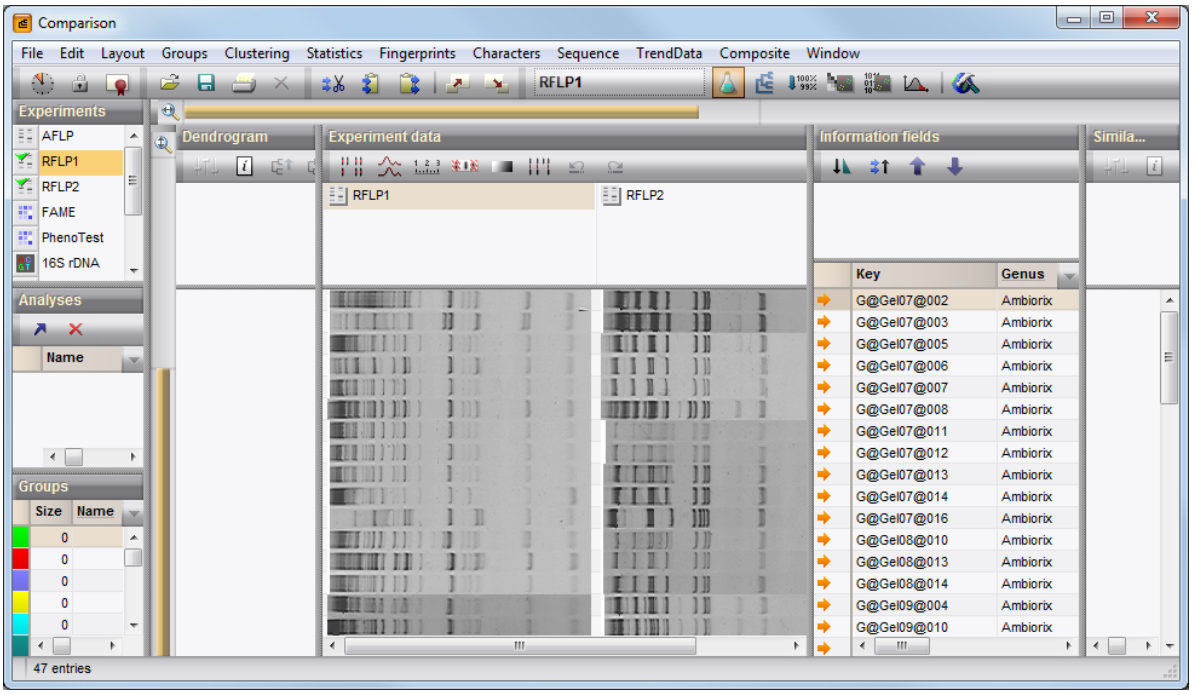


Figure 3.1.3: The *Comparison window*: Two fingerprint types are shown.

4.5 If there is not enough space to show both images at the same time, scroll through the data panel, or use the zoom slider.

4.6 Drag the separator line between the experiments to the left or to the right to adjust the horizontal space for a particular experiment.

4.7 Select experiment type **RFLP1** in the *Comparison window* by selecting its name in the *Experiments panel* or click on the RFLP1 image in the *Experiment data panel* (if shown).

Functions like clustering, PCA, and band matching, as well as layout functions, apply to the currently selected experiment type. When performing any of these functions, be sure the correct experiment is selected!

4.8 Press next to **PhenoTest** in the *Experiments panel*.

The colors defined for the **PhenoTest** experiment type are shown in the *Experiment data panel* (see Figure 3.1.4).

4.9 Press next to **16S rDNA** in the *Experiments panel*.

The sequences are displayed in the *Experiment data panel* (see Figure 3.1.4).

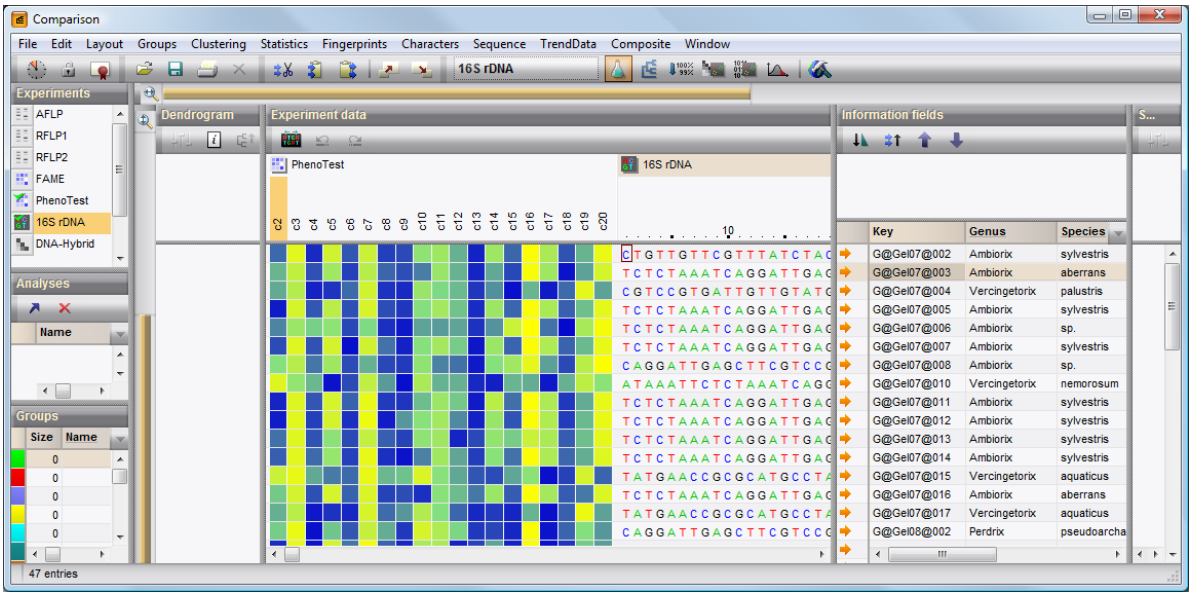


Figure 3.1.4: The *Comparison window*: Character type & Sequence type.

3.1.4.2 Add and remove entries

4.10 Unselect all entries by pressing the **F4** key.

4.11 Select a few entries in the *Information fields panel* of the *Comparison window* (use the **Ctrl** and **Shift**–keys to select/unselect entries).

4.12 Press and confirm the action. The selected entries are removed from the comparison.

4.13 Press . The selected entries are pasted back into the comparison at the position of the selection bar.

4.14 Select *File* > *Save* (, **Ctrl+S**) to save the comparison and enter "All" as the name. Press <***OK***>.

4.15 Select *File* > *Exit* to close the comparison.

Comparison **All** is now listed in the *Comparisons panel* of the *BioNumerics main window*.

3.1.4.3 Create groups

An important display function in the *Comparison window* is the creation of groups. Groups are subsets of comparison entries that can be defined from clusters, from database fields, or from any subdivision the user

desires.

As an example we will use the **Genus** database field to assign groups in our comparison.

4.16 In the **DemoBase Connected** database, double–click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non–STANDARD entries (see 3.1.3).

4.17 Press the **F4** key to clear any selected entries. Right–click on the database field name **Genus** in the *Information fields panel* and select *Create groups from database field*.

4.18 Select the order in which groups are created (i.e. by size, alphabetically, or by position in the comparison) and press <***OK***> to create the groups.

Every genus is now assigned to a unique group. They appear in the *Groups panel* along with their sizes and names (see Figure 3.1.5).

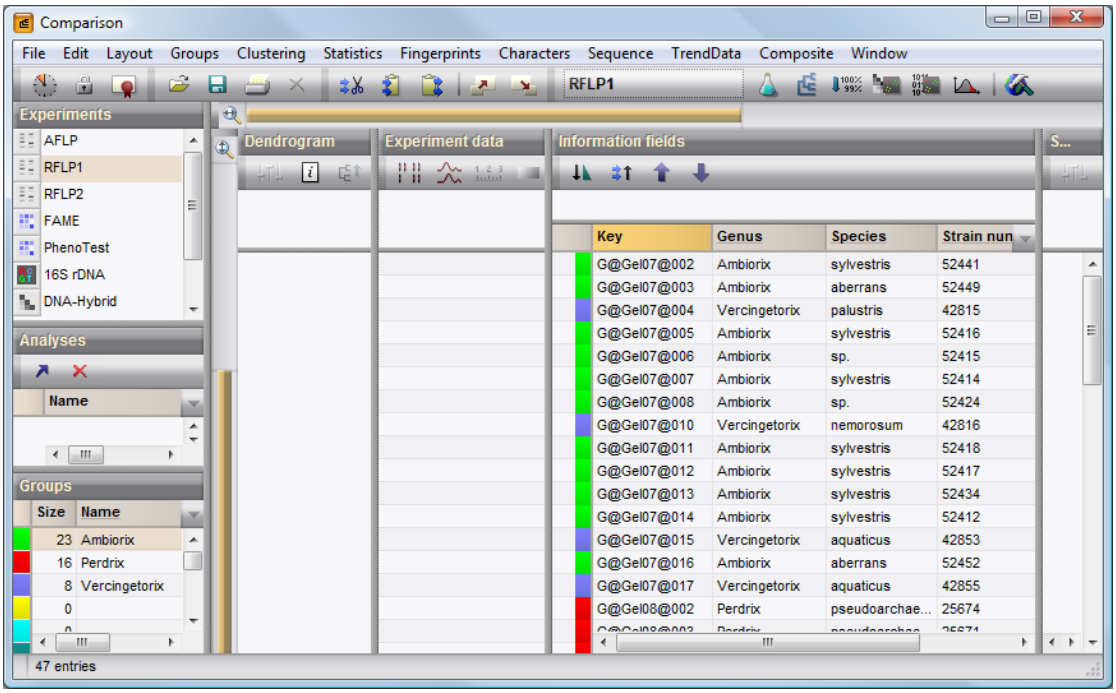


Figure 3.1.5: The *Comparison window* with groups defined.

4.19 Save the comparison and close the *Comparison window*.

Chapter 3.2

Clustering fingerprint data

3.2.1 Comparison window

1.1 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non-STANDARD entries (see 3.1.3).

1.2 Press next to **RFLP1**. The fingerprints for **RFLP1** are shown in the *Experiment data panel* (see Figure 3.1.3).

1.3 Select *Fingerprints* > *Settings* > *Show metrics scale* ( ) to display the metric (e.g. molecular weight) scale of the selected fingerprint type.

1.4 Press to show the band positions in the *Experiment data panel*.

3.2.2 Clustering fingerprint data

Cluster analysis is a two-step process. First, all pairwise similarity values are calculated with a **similarity coefficient**. Then, the resulting similarity matrix is converted into a dendrogram with a **clustering algorithm**. Although in practice these steps are performed together, they each require their own comparison settings.

2.1 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

The *Comparison settings wizard* allows you to specify the settings related to the similarity coefficient for calculation of the similarity matrix and the clustering method to be applied. The first step deals with the similarity coefficient (see Figure 3.2.1).

2.2 Select *Dice* from the list.

Additional settings are listed in the right panel.

2.3 Enter a *Optimization* of 0.50%, and a *Band matching Tolerance* of 0.50%. Leave all the other settings to 0% (see Figure 3.2.1).

The *Optimization* setting limits the amount of movement for each fingerprint as a whole. The *Band matching Tolerance setting* limits the amount of movement for each band.

2.4 Press <***Next***>.

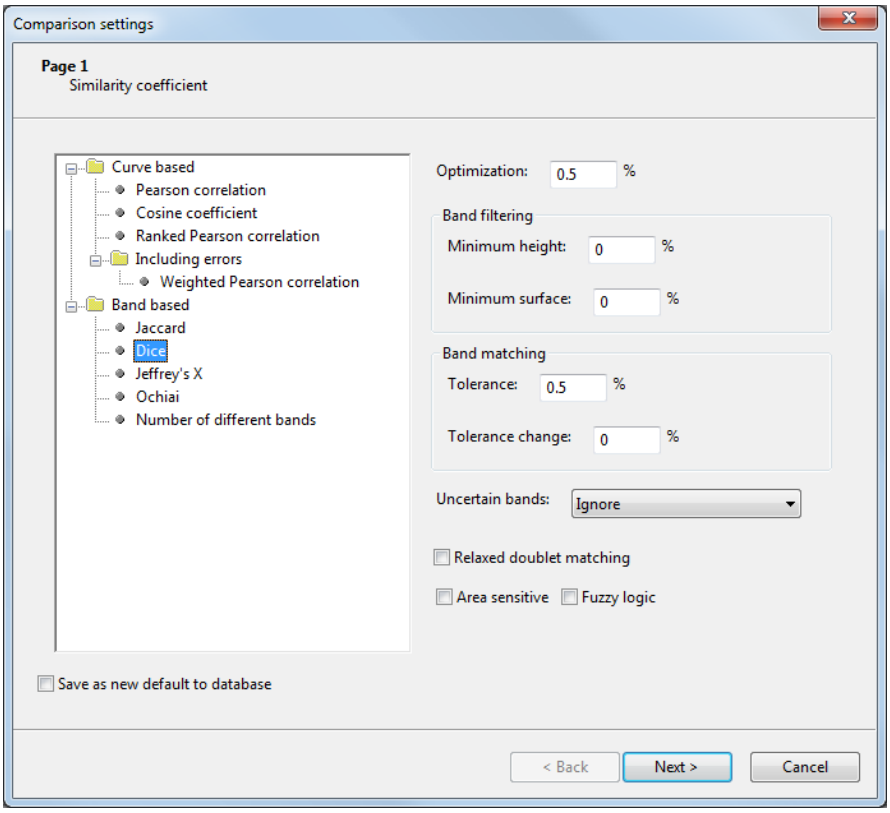


Figure 3.2.1: Select similarity coefficient.

In step 2 of the *Comparison settings wizard*, the options related to the clustering algorithms are grouped (see Figure 3.2.2). Under *Method*, the clustering algorithm to be applied on the similarity matrix can be selected. A *Dendrogram name* can be entered in the corresponding text box. By default, the name of the experiment type will be used.

2.5 Select *UPGMA*, check *Calculate error flags* and select *Cophenetic correlation* from the *Branch quality* list (see Figure 3.2.2).

If *Calculate error flags* is checked, the program will calculate the standard deviations associated with each cluster. The *Cophenetic Correlation* is another parameter that expresses the consistency of a cluster. This method calculates the correlation between the dendrogram-derived similarities and the matrix similarities. The value is calculated for each cluster thus estimating the faithfulness of each sub-cluster of the dendrogram.

2.6 Press <***Next***> again in the *Comparison settings wizard* to start the cluster analysis.

During the calculations, the program shows the progress in the *Comparison window*'s caption (as a percentage), and there is a green progress bar in the bottom of the window.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The cluster analysis is listed in the *Analyses panel* of the *Comparison window* (see Figure 3.2.3).

The *Cophenetic correlation* is shown at each branch, together with a colored dot, of which the color ranges between green-yellow-orange-red according to decreasing cophenetic correlation. This makes it easy to detect reliable and unreliable clusters at a glance.

Blue bars are also shown at each node, corresponding to the *Standard deviation* of values in that region of the similarity matrix. The average and the standard deviation of similarity values for the selected node are

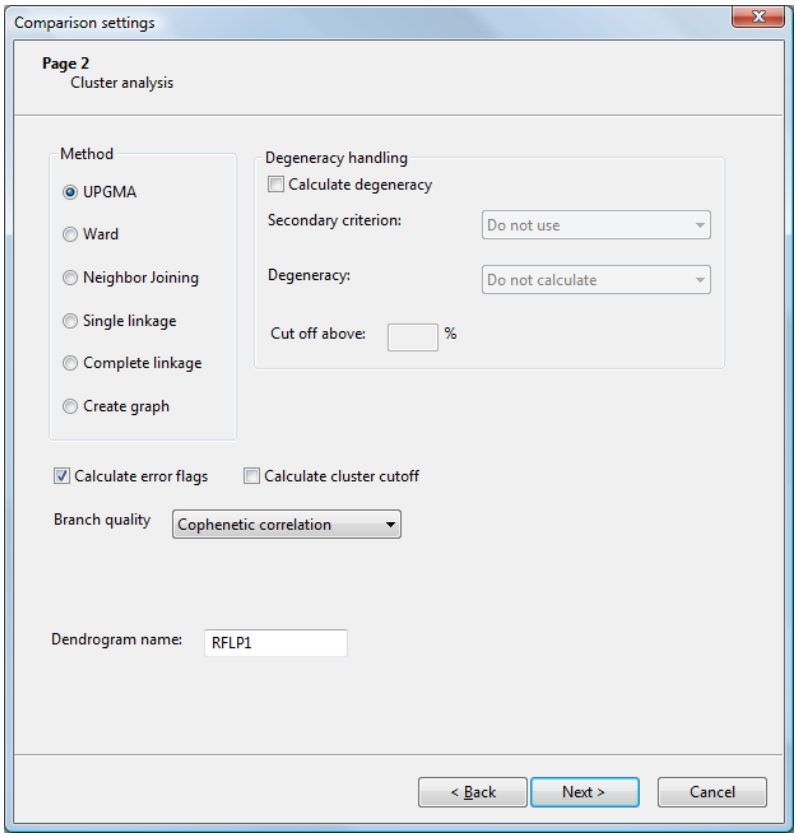


Figure 3.2.2: Select clustering algorithm.

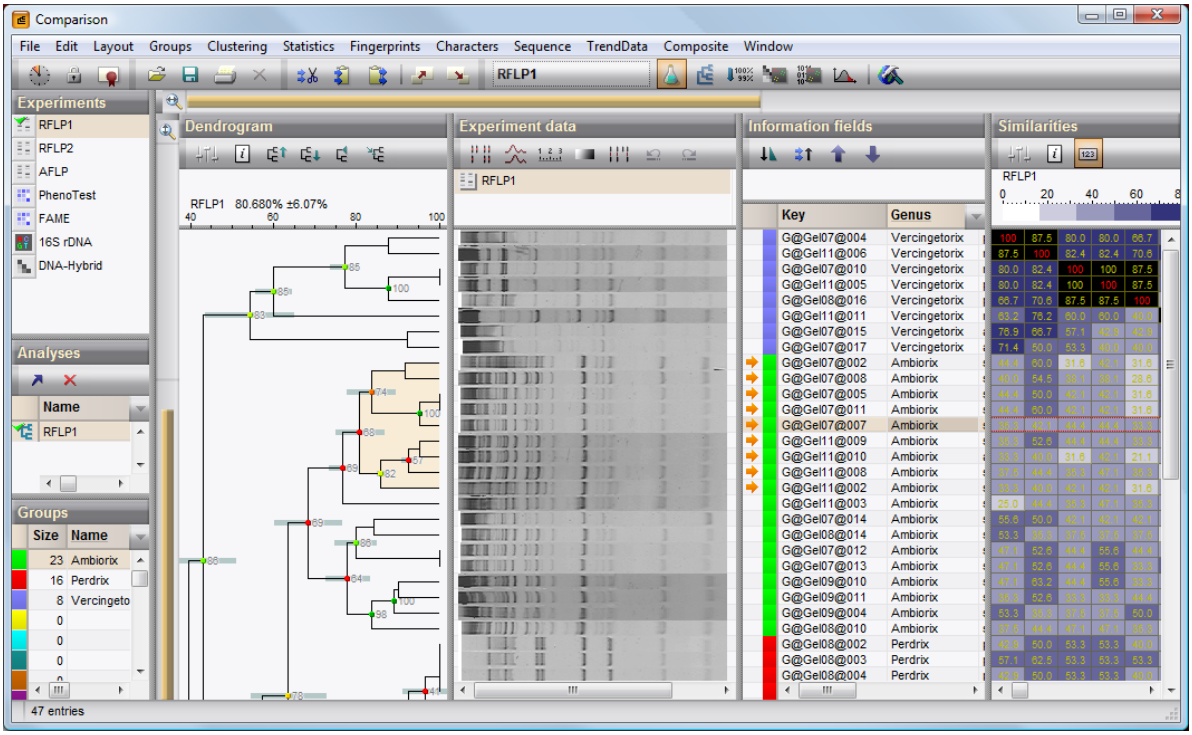


Figure 3.2.3: The *Comparison window*.

shown above the dendrogram.

2.7 Left-click on the dendrogram to place the cursor on any node or tip (where a branch ends in an individual entry).

2.8 To select entries in a cluster, click on the node of the cluster while holding the Ctrl-button.

2.9 Press to remove the selected entries from the cluster analysis. Confirm the action. The dendrogram is automatically updated.

2.10 Select *Edit* > *Paste selection* or . The cluster analysis is recalculated automatically, and the selected entries are placed back in the dendrogram.

A branch can be moved up or down to improve the layout of a dendrogram:

2.11 Click the branch which you want to move up in the dendrogram and select *Clustering* > *Move branch up* or press the button in the *Dendrogram panel*.

2.12 Click the branch which you want to move down in the dendrogram and select *Clustering* > *Move branch down* or press the button in the *Dendrogram panel*.

To simplify the representation of large and complex dendrograms, it is possible to simplify branches by abridging them as a triangle.

2.13 Select a cluster of closely related entries and select *Clustering* > *Collapse/expand branch* or press the button in the *Dendrogram panel*. Repeat this action to undo the abridge operation.

2.14 If no groups are defined in the *Comparison window*, right-click on the field name **Genus** in the *Information fields panel*, select *Create groups from database field* and confirm.

2.15 Select *Clustering* > *Dendrogram display settings* or press the button in the *Dendrogram panel*.

This pops up the *Dendrogram display settings dialog box*.

2.16 Uncheck *Show error flags*, uncheck *Show branch quality*, and enable *Show group colors*. Press <***OK***>.

The dendrogram branches are now colored according to the group colors (see Figure 3.2.4).

2.17 Select *Clustering* > *Show information* or press in the *Dendrogram panel*. This pops up a *Report window* with the comparison settings. Close the *Report window*.

2.18 Save the comparison with the dendrogram by selecting *File* > *Save* or pressing .

3.2.3 Matrix display functions

The similarity values in the *Similarities panel* are represented by shades of blue.

3.1 To show the values in the matrix, select in the toolbar of the *Similarities panel*.

3.2 To view a pairwise comparison, double-click on the appropriate cell in the matrix.

3.2.4 Printing a cluster analysis

BioNumerics can export the cluster analysis as it appears in the *Comparison window*.

4.1 Select *File* > *Print preview...* (, **Ctrl+P**).

4.2 To scan through the pages that will be printed out, use *Edit* > *Previous page* (, **Pge Up**) and *Edit* > *Next page* (, **Pge down**).

4.3 To zoom in or out, use *Edit* > *Zoom in* (, **Ctrl+Pge Up**) and *Edit* > *Zoom out* (, **Ctrl+Pge Down**) or use the zoom slider.

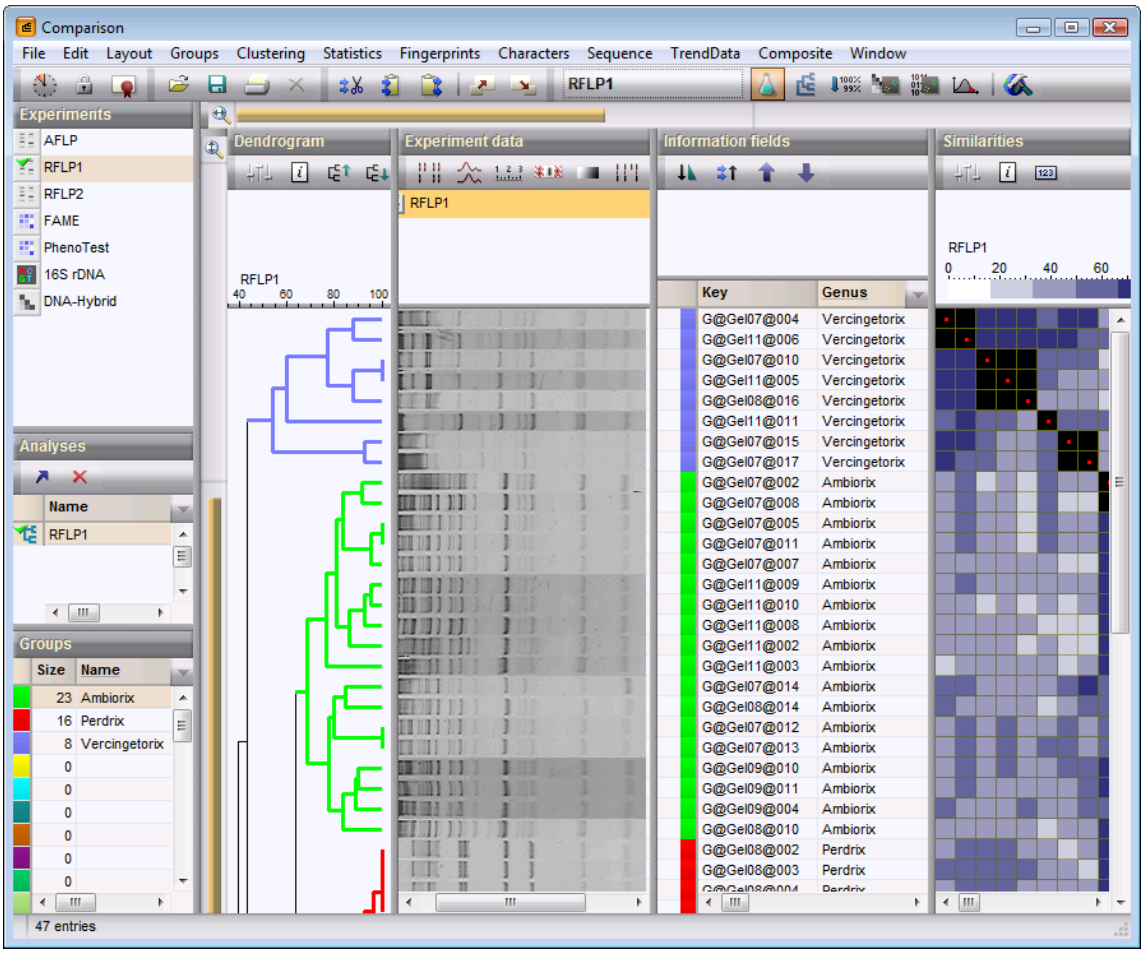


Figure 3.2.4: Show group colors on dendrogram.

4.4 To enlarge or reduce the whole image, use *Layout* > *Enlarge image size* ( ) or *Layout* > *Reduce image size* ( ).

4.5 If a similarity matrix is available, it can be included with *Layout* > *Show similarity matrix* ( ).

4.6 On top of the page, there are a number of small yellow slider bars, which can be moved.

4.7 To preview and print the image in full color select *Layout* > *Use colors* ( ).

4.8 Export the image to the clipboard with *File* > *Copy page to clipboard* ( ) and selecting an appropriate format.

4.9 If a printer is available, use *File* > *Print all pages* ( ) or *File* > *Print this page* ( ) to print one or all pages.

4.10 Select *File* > *Exit* to close the *Comparison print preview window*.

3.2.5 Additional practice

5.1 In the **E. coli** database - created in 1.2.1 - create a comparison of all the database entries except **REF**.

5.2 Create groups from the **Source** database field.

5.3 Show the **PFGE-XbaI** patterns and perform a cluster analysis using *Dice* and *UPGMA*.

Chapter 3.3

Clustering character data

3.3.1 Comparison window

1.1 In the **DemoBase Connected** database, double-click on **PhenoTest** in the *Experiments panel*.

Twenty characters are listed, with a character range of 0 to 3. The color scale for each character ranges from yellow to blue, with a green transition color.

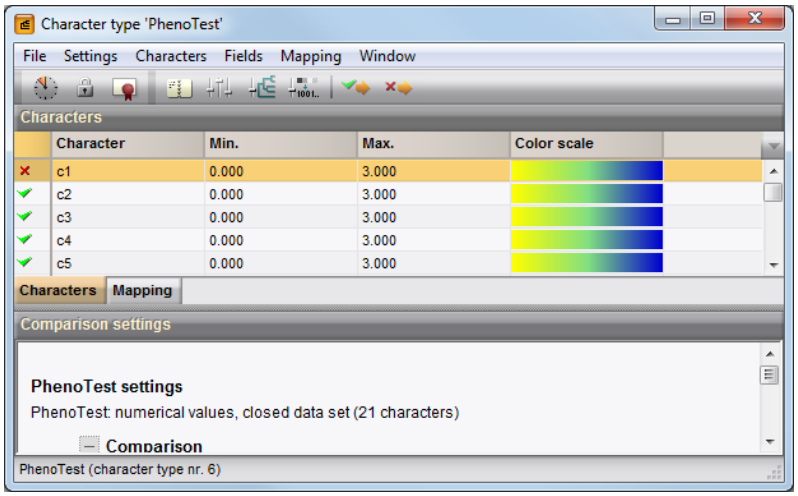


Figure 3.3.1: The *Character type window*.

1.2 Close the *Character type window*.

1.3 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non-STANDARD entries (see 3.1.3).

1.4 Select **PhenoTest** in the *Experiments panel* and press the icon next to **PhenoTest**.

The character values are displayed as colors according to the color scale defined for each character in the *Character type window* (see Figure 3.3.1).

1.5 Select *Characters* > *Show values* ( ) to show the character values for all entries.

1.6 The character values can be displayed as colors again with *Characters* > *Show colors* ( ).

1.7 Select a character in the header of the *Experiment data panel* (e.g. **c5**).

1.8 Select *Characters* > *Sort by character value* ( ) in the *Experiment data panel*.

The entries are now sorted by increasing value of the selected character (see Figure 3.3.2).

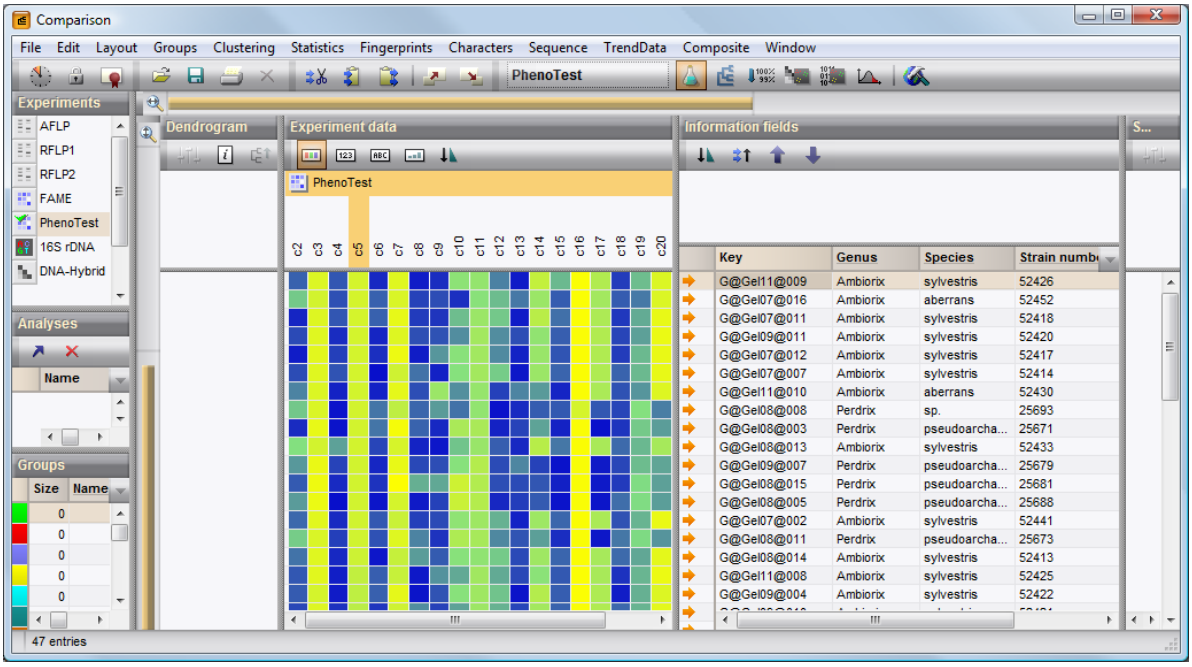


Figure 3.3.2: The *Comparison window*: character type experiment.

3.3.2 Clustering character data

Cluster analysis is a two-step process. First, all pairwise similarity values are calculated with a **similarity coefficient**. Then, the resulting similarity matrix is converted into a dendrogram with a **clustering algorithm**. Although in practice these steps are performed together, they each require their own settings.

2.1 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

The *Comparison settings wizard* allows you to specify the settings related to the similarity coefficient for calculation of the similarity matrix and the clustering method to be applied. The first step deals with the similarity coefficient (see Figure 3.3.3).

2.2 Select *Pearson correlation* from the list and press <***Next***> (see Figure 3.3.3).

In step 2 of the *Comparison settings wizard*, the options related to the clustering algorithms are grouped (see Figure 3.2.2). Under *Method*, the clustering algorithm to be applied on the similarity matrix can be selected. A *Dendrogram name* can be entered in the corresponding text box. By default, the name of the experiment type will be used.

2.3 Select *UPGMA*, check *Calculate error flags* and select *Cophenetic correlation* from the *Branch quality* list (see Figure 3.2.2).

If *Calculate error flags* is checked, the program will calculate the standard deviations associated with each cluster. The *Cophenetic Correlation* is another parameter that expresses the consistency of a cluster. This method calculates the correlation between the dendrogram-derived similarities and the matrix similarities.

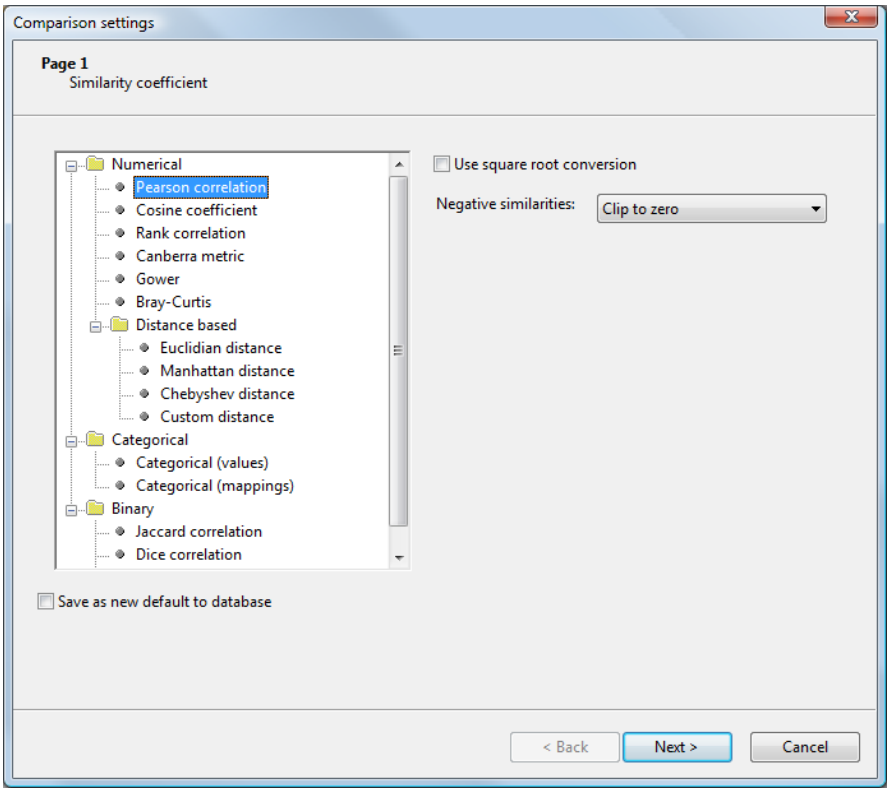


Figure 3.3.3: Select similarity coefficient.

The value is calculated for each cluster thus estimating the faithfulness of each sub-cluster.

2.4 Press <***Next***> again in the *Comparison settings wizard* to start the cluster analysis.

During the calculations, the program shows the progress in the *Comparison window*'s caption (as a percentage), and there is a green progress bar in the bottom of the window.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The cluster analysis is listed in the *Analyses panel* of the *Comparison window* (see Figure 3.3.4).

The *Cophenetic correlation* is shown at each branch, together with a colored dot, of which the color ranges between green-yellow-orange-red according to decreasing cophenetic correlation. This makes it easy to detect reliable and unreliable clusters at a glance.

Blue bars are also shown at each node, corresponding to the *Standard deviation* of values in that region of the similarity matrix. The average and the standard deviation of similarity values for the selected node are shown above the dendrogram.

2.5 Left-click on the dendrogram to place the cursor on any node or tip (where a branch ends in an individual entry).

2.6 To select entries in a cluster, click on the node of the cluster while holding the Ctrl-button.

2.7 Press to remove the selected entries from the cluster analysis. Confirm the action. The dendrogram is updated.

2.8 Select *Edit* > *Paste selection* or . The cluster analysis is recalculated automatically, and the selected entries are placed back in the dendrogram.

A branch can be moved up or down to improve the layout of a dendrogram:

2.9 Click the branch which you want to move up in the dendrogram and select *Clustering* > *Move branch up* or press the button in the *Dendrogram panel*.

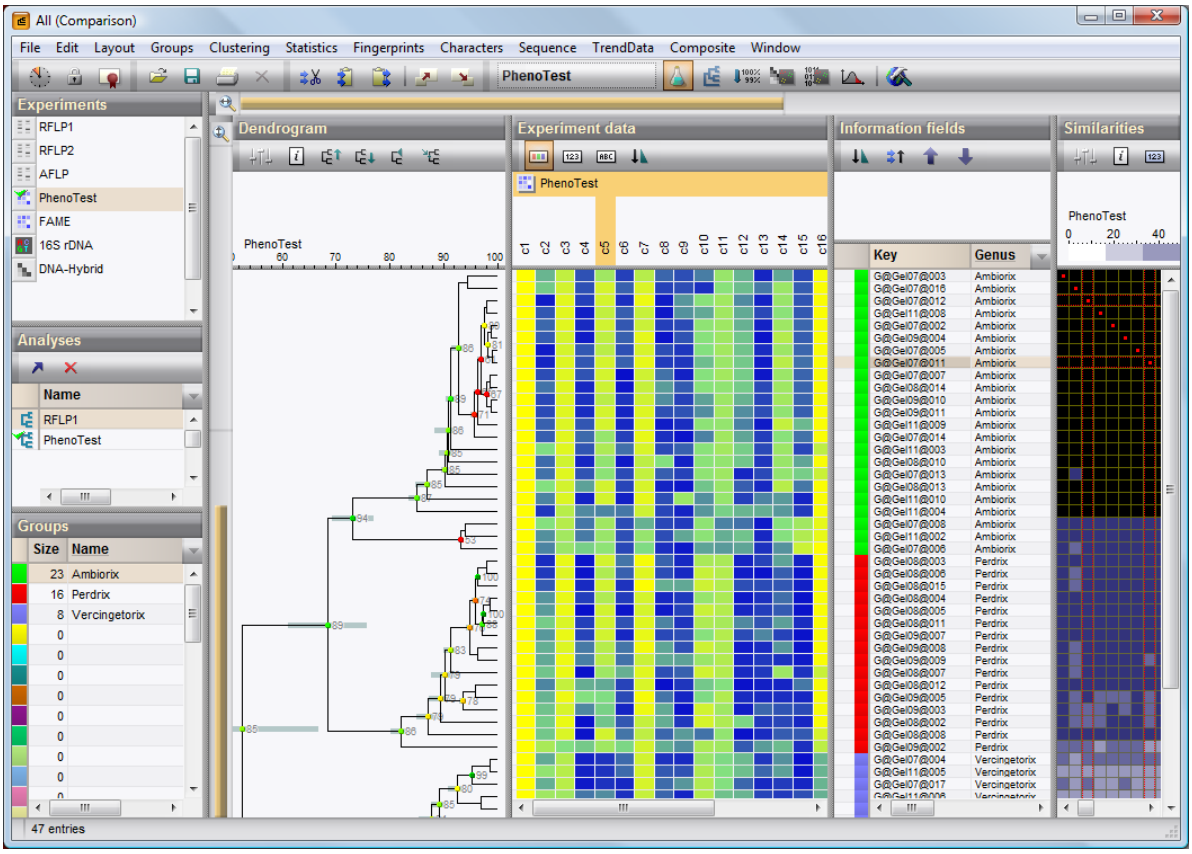


Figure 3.3.4: The *Comparison window*.

2.10 Click the branch which you want to move down in the dendrogram and select *Clustering* > *Move branch down* or press the button in the *Dendrogram panel*.

To simplify the representation of large and complex dendrograms, it is possible to simplify branches by abridging them as a triangle.

2.11 Select a cluster of closely related entries and select *Clustering* > *Collapse/expand branch* or press the button in the *Dendrogram panel*. Repeat this action to undo the abridge operation.

2.12 If no groups are defined in the *Comparison window*, right-click on the field name **Genus** in the *Information fields panel*, select *Create groups from database field* and confirm.

2.13 Select *Clustering* > *Dendrogram display settings* or press the button in the *Dendrogram panel*.

This pops up the *Dendrogram display settings dialog box*.

2.14 Uncheck *Show error flags*, uncheck *Show branch quality*, and enable *Show group colors*. Press <***OK***>.

The dendrogram branches are now colored according to the group colors.

2.15 Select *Clustering* > *Show information* or press in the *Dendrogram panel*. This pops up a *Report window* with the comparison settings. Close the *Report window*.

2.16 Save the comparison with the dendrogram, by selecting *File* > *Save* or pressing .

More information on the tools that are available in the *Comparison window* can be found in 3.2.3 and 3.2.4, and in the BioNumerics manual.

Chapter 3.4

Sequence alignment and clustering

3.4.1 Introduction

Sequence alignment is inseparable from cluster analysis. In this Chapter we will perform a pairwise cluster analysis, then a multiple alignment, followed by a global cluster analysis in both the *Comparison* and the *Alignment window*.

3.4.2 Comparison window

3.4.2.1 Pairwise sequence cluster analysis

First we will calculate the similarities between all pairs of sequences based on pairwise alignments.

2.1 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non–STANDARD entries (see 3.1.3).

2.2 Press next to the sequence type **16S rDNA** in the *Experiments panel* to display the sequences.

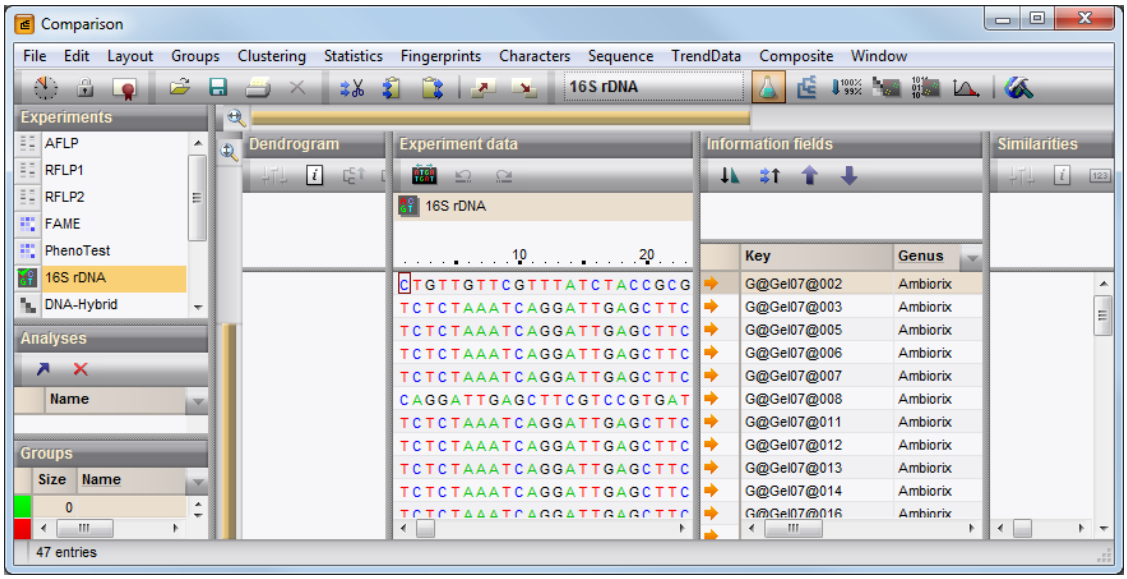


Figure 3.4.1: The *Comparison window*: Sequence type experiment (initial view).

Initially, the sequences are not aligned and no similarity matrix exists (see Figure 3.4.1).

2.3 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

The *Comparison settings wizard* appears. The settings are shown in the right panel of the dialog box and depend on the algorithm selected in the left panel (see Figure 3.4.2).

2.4 Select *Standard* under *Pairwise alignment*, leave the other settings unaltered and press <***Next***>.

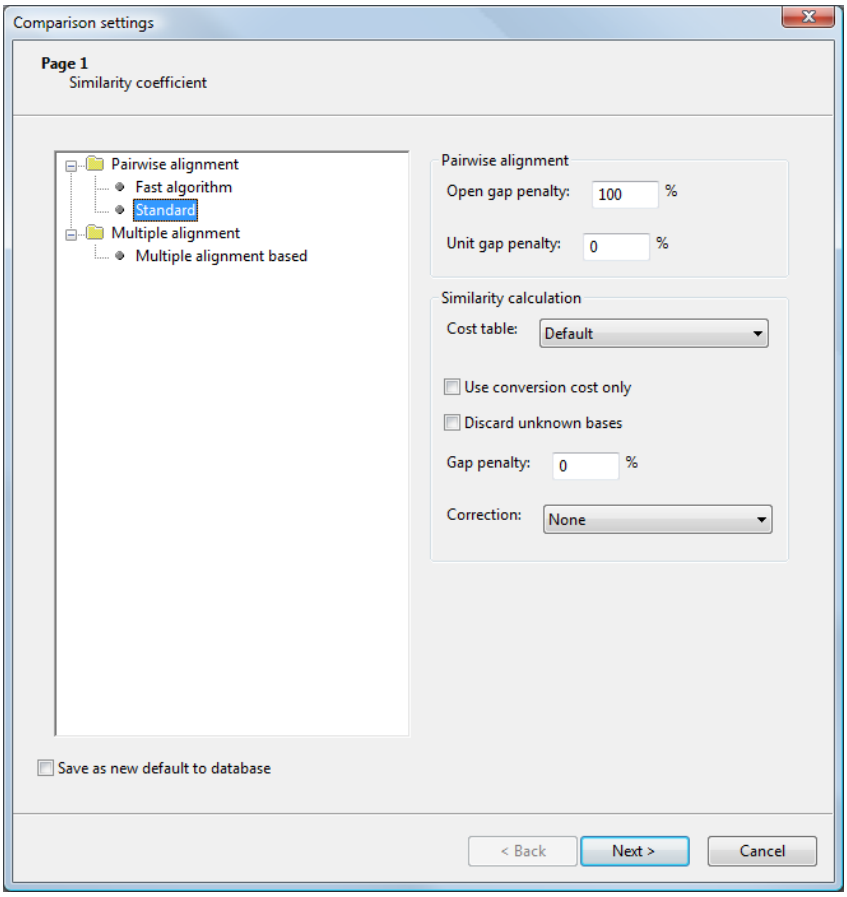


Figure 3.4.2: Select similarity coefficient.

In step 2 of the *Comparison settings wizard*, the options related to the clustering algorithms are grouped (see Figure 3.2.2). Under *Method*, the clustering algorithm to be applied on the similarity matrix can be selected. A *Dendrogram name* can be entered in the corresponding text box. By default, the name of the experiment type will be used.

2.5 Select *UPGMA* and press <***Next***>.

During the calculations, the program shows the progress in the *Comparison window*'s caption (as a percentage), and there is a green progress bar in the bottom of the window.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The cluster analysis is listed in the *Analyses panel* of the *Comparison window*. The sequences are still unaligned.

3.4.2.2 Multiple sequence alignment

The multiple alignment is derived from the dendrogram created by the pairwise clustering. Each node in the dendrogram represents a consensus sequence for that cluster, created during the construction of the

dendrogram. The root node represents the consensus sequence for the entire dendrogram. This global consensus is used to align all of the sequences in the comparison simultaneously.

2.6 Select *Sequence* > *Multiple alignment* or .

2.7 Accept the default settings and press <***OK***> to start the multiple alignment.

When the calculations are done, the sequences are aligned in the *Experiment data panel* (see Figure 3.4.3).

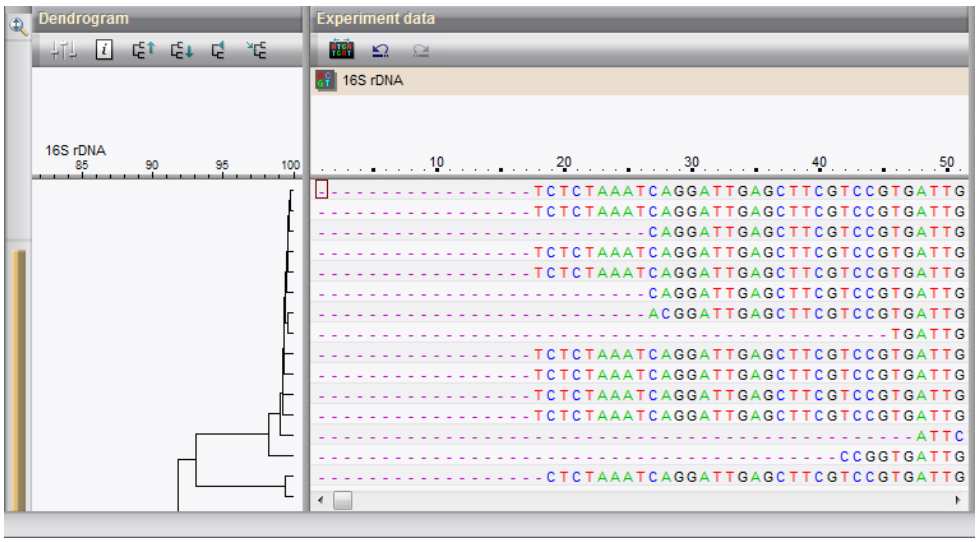


Figure 3.4.3: Multiple sequence alignment.

2.8 Select *Sequence* > *Block type* > *Neighbor blocks* to show the *Neighbor match* representation. Bases that differ from their neighbors are highlighted (see Figure 3.4.4).

Next, we are going to create a consensus sequence based on all sequences in the *Comparison window*.

2.9 For this exercise, select the root and *Sequence* > *Create consensus of branch*.

A dialog box prompts "Enter minimum consensus percentage". If a minimum percentage of 50 is specified, a base at a given position will only be shown in the consensus sequence if at least 50% of the sequences have that base at the given position.

2.10 Enter a minimum consensus percentage of "50" and press <***OK***>.

The global consensus sequence is shown above the sequences in the *Experiment data panel*. For every position at which at least 50% of the aligned sequences agree, that base appears in the consensus sequence. Otherwise, the position is labeled N.

2.11 Select *Sequence* > *Block type* > *Consensus blocks* to highlight the bases that are identical to the consensus sequence (see Figure 3.4.4).

2.12 Select *Sequence* > *Block type* > *Consensus difference* to show only the bases that differ from the consensus sequence (see Figure 3.4.4).

2.13 A consensus sequence can be copied to the clipboard with *Sequence* > *Copy consensus to clipboard*.

2.14 Save the comparison by selecting *File* > *Save* or pressing .

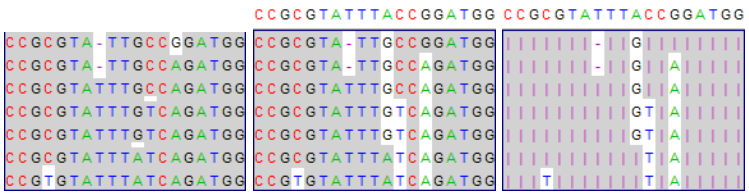


Figure 3.4.4: Neighbor match; Consensus match; Consensus difference.

3.4.2.3 Sequence cluster analysis based on multiple alignment

We can now cluster the sequences based on the multiple alignment. Since the multiple alignment differs from the initial pairwise alignments, the resulting similarity matrix will differ somewhat from that of the pairwise clustering.

2.15 To calculate a global clustering based on the alignment that is present in the *Experiment data panel*, select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)....*

The *Comparison settings wizard* appears (see Figure 3.4.2). The settings are shown in the right panel of the dialog box and depend on the algorithm selected in the left panel.

2.16 Select the *Multiple alignment based* option under *Multiple alignment* in the left panel.

2.17 Use the *Default* cost table, check *Discard unknown bases*, select a *Gap penalty* of 0, apply no correction and leave *Use active zones only* unchecked. Press <***Next***>.

2.18 Select *Neighbor Joining* as the clustering method in the next step and press <***Next***> to calculate the global cluster analysis (see Figure 3.4.5).

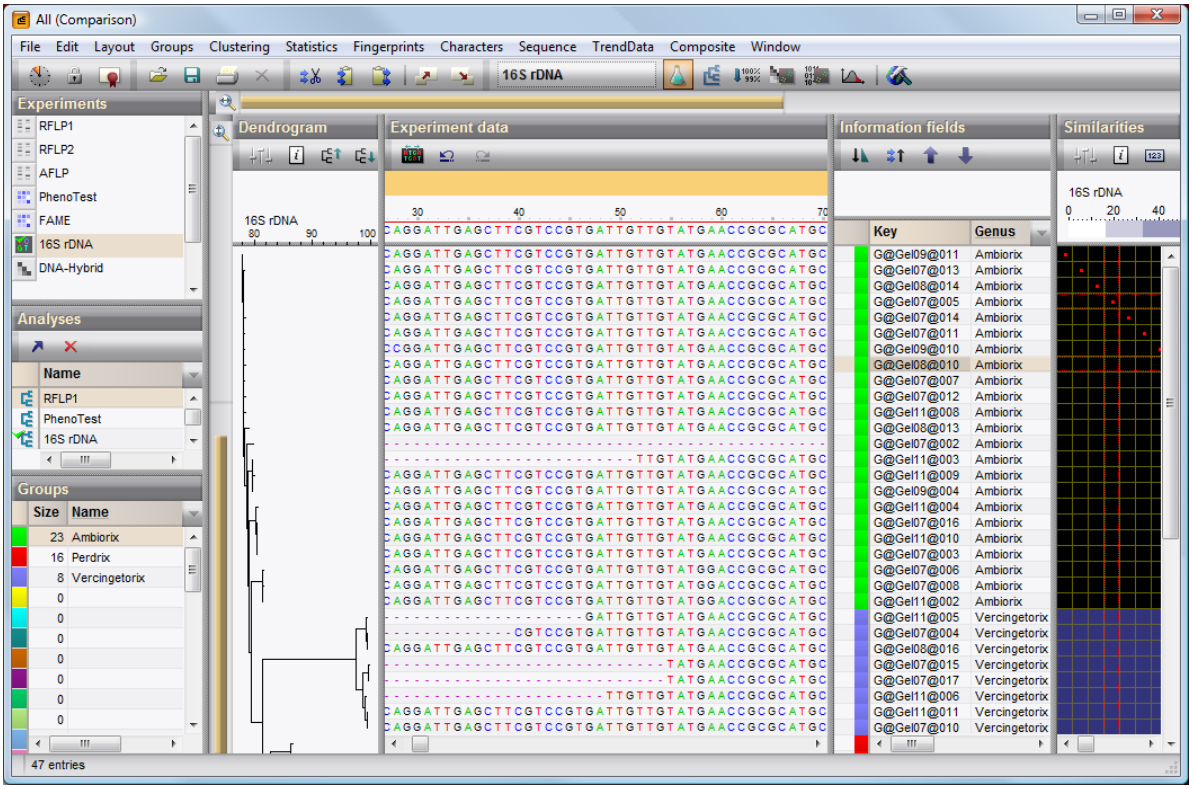


Figure 3.4.5: The *Comparison window*.

3.4.2.4 Exporting a multiple alignment

2.19 To export the sequences as a text file select *File* > *Export* > *Export sequences (tabular)*.

2.20 Use the *File* > *Export* > *Export sequences (formatted)* to export the sequences in a more advanced way.

2.21 Save and close the *Comparison window*.

3.4.3 Alignment window

The *Alignment window* is a convenient tool for the calculation of multiple sequence alignments, subsequence searches and mutation analysis. In this Section, the influenza dataset in the database **SeqAssembly** will be used to illustrate some basic features in the *Alignment window*.

3.1 Open the **SeqAssembly** database (see 2.3.4 for the creation of the database and import of the files).

3.2 Make sure the *Alignments panel* is displayed in the *BioNumerics main window* (see Figure 3.4.6).

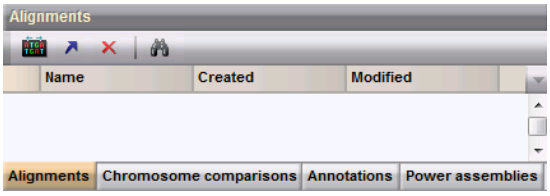


Figure 3.4.6: The *Alignments panel*.

3.3 Select *Comparison* > *Alignments* > *Create new* or press the button in the *Alignments panel*.

3.4 Enter a name, for example **MyAlign** and press <***OK***>.

3.5 Press **Ctrl+A** to select all entries in the database and press the button in the *Alignments panel* to open the project with the selected entries. Press <***OK***>.

3.4.3.1 Sequence display

3.6 In the *Alignment window*, press the button to load the curves in the *Sequence display 2 panel*.

3.7 Select *Options* > *Sequence display 2*, check *Show multi line curves*, *Show sequence*, and *Use sequence color codes*. Press <***OK***>.

The curves are displayed on different lines. The consensus sequence is shown on top of the chromatograms.

3.8 Press to show the translated amino acid sequences.

3.4.3.2 Alignment and clustering

3.9 Select *Alignment* > *Calculate* > *Multiple alignment* or press the button.

The settings for the successive pairwise and multiple alignment steps are shown in the new window.

3.10 Select <***Defaults***> and press <***OK***>.

The dendrogram and similarity matrix are still based on the pairwise similarity values.

3.11 To calculate the clustering based on the multiple alignment press the button.

3.12 Select *Neighbor Joining*, leave all other settings unaltered, and press <***OK***>.

3.4.3.3 Mutation and SNP analysis

Before we can search for mutations, we first need to calculate a consensus sequence.

3.13 Make sure all entries are selected (**Ctrl+A**) and select *Alignment* > *Consensus* > *Create from selected entries* or press .

3.14 In the dialog box, leave the first setting unaltered, enter **20** as the "Minimal fraction of a specific nucleotide ..." and press <***OK***>.

The consensus is displayed in the header of the *Sequence display 1 panel*.

3.15 Select *Mutations* > *Search* or press the button in the *Mutation listing panel*.

3.16 Leave all settings at their defaults and press <***Find***> to start the mutation search.

The results are displayed in the *Mutation listing panel*.

3.17 Click on any of the mutations listed in the panel.

The cursor will jump to the corresponding position on the alignment and the curves.

3.18 Save the project and close the *Alignment window*.

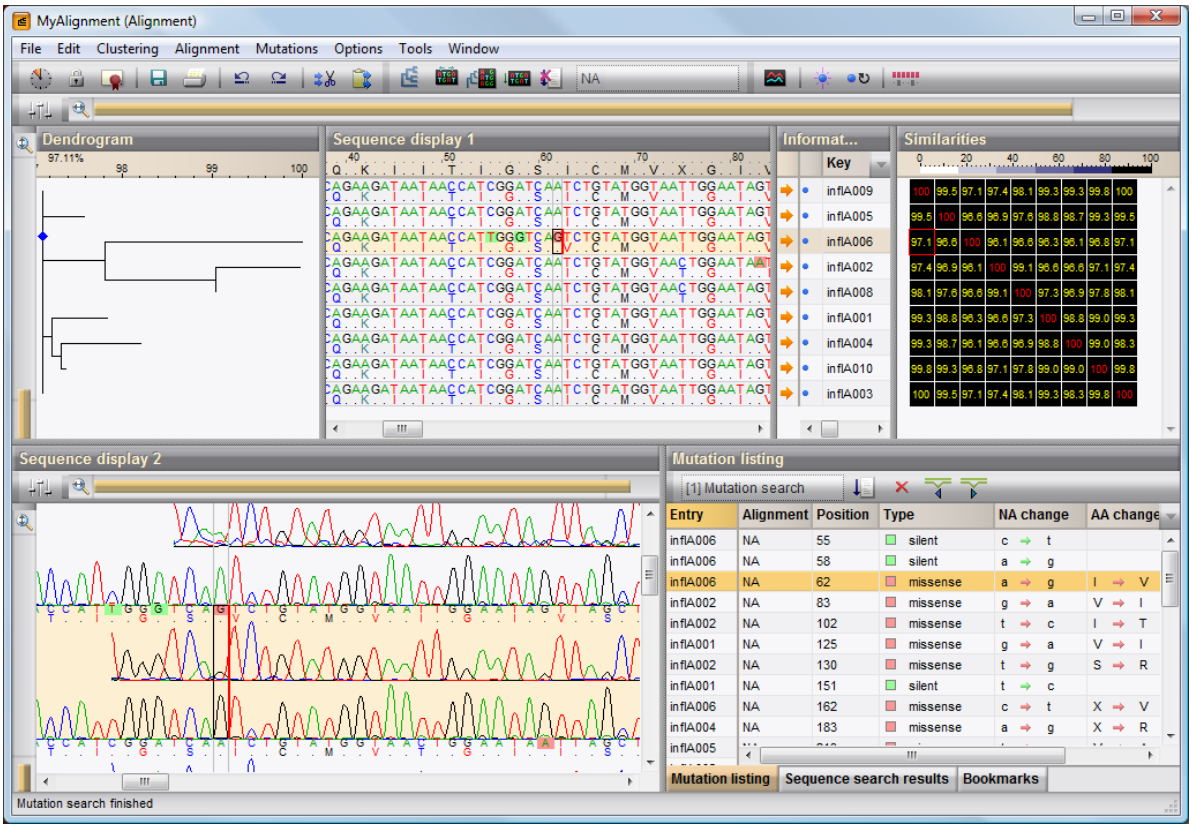


Figure 3.4.7: The *Alignment window*.

Chapter 3.5

Band matching tables

3.5.1 Creating a band matching table

Fingerprint patterns do not have well-defined characters. Band positions vary continuously, although they do tend to fall into categories, or *band classes*. BioNumerics allows you to formally define band position classes, thereby creating a band matching table. This in turn allows you to apply more sophisticated analytical tools, such as polymorphism analysis and principal components analysis.

3.5.1.1 Creating a composite data set

As will be described in the next Chapter, a composite data set can be used to combine two or more experiments. A composite data set can also be used to convert fingerprint band classes into a band matching table. As an exercise we will define a composite data set containing the fingerprint experiment **RFLP1**.

1.1 In the **DemoBase Connected** database, press from the *Experiments panel toolbar* and select *New composite data set*.

1.2 Enter **RFLP1-table** and press <***OK***>.

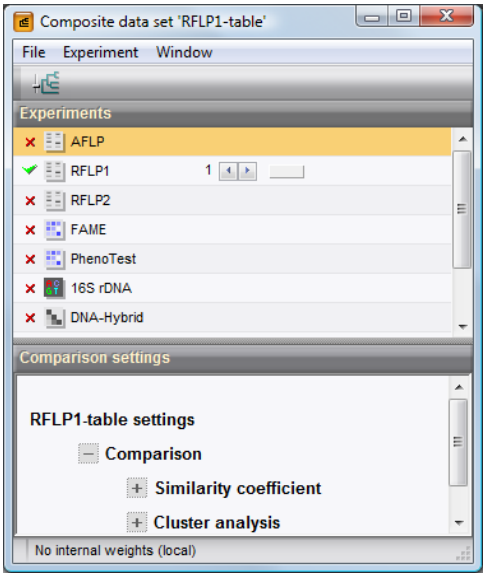


Figure 3.5.1: The *Composite data set window*.

1.3 Select **RFLP1** in the *Composite data set window* and select *Experiment* > *Use in composite data set*.

The **RFLP1** experiment is checked with a green V-sign (see Figure 3.5.1).

1.4 Close the *Composite data set window*.

RFLP1-table is shown in the *Experiments panel* of the *BioNumerics main window* (see Figure 3.5.2).

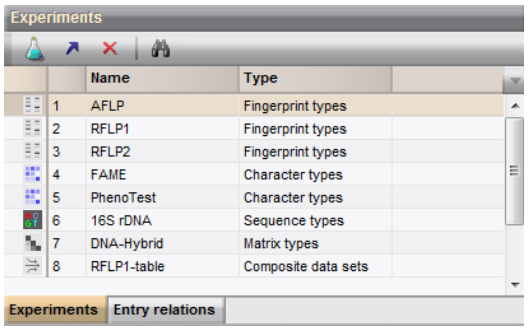


Figure 3.5.2: The *Experiments panel*.

3.5.1.2 Creating band classes

Band classes are position categories to which individual bands are assigned. Just as the position tolerance and optimization settings determine how a pair of fingerprint patterns are aligned, they also determine how band classes are created for a whole comparison. The result is a band table with defined columns corresponding to band class positions.

1.5 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non-STANDARD entries (see 3.1.3).

1.6 In the *Comparison window*, select **RFLP1** in the *Experiments panel* and press .

1.7 Select *Fingerprints* > *Perform band matching* or press the button.

1.8 Select *Find classes on all entries* and press <***OK***>.

The software defines the band classes for **RFLP1** and assigns each band to a class. The classes are shown as blue lines (see Figure 3.5.3).

1.9 Use the zoom sliders to obtain the best view of the band classes.

All band classes are labeled with a band class label. If a band class is selected, its label is highlighted. If a regression curve is calculated for the reference system, the metric positions of the band classes are displayed in the label.

3.5.1.3 Displaying the band matching table

The band classes are shown as blue lines crossing the fingerprints. To analyze them further, the band classes must be displayed as a band table, using a composite data set.

1.10 Show the band matching table by pressing next to **RFLP1-table**.

Each cell in the table represents a band's presence or absence.

1.11 To see the band class labels completely, drag the separator line between the table and the labels downwards.

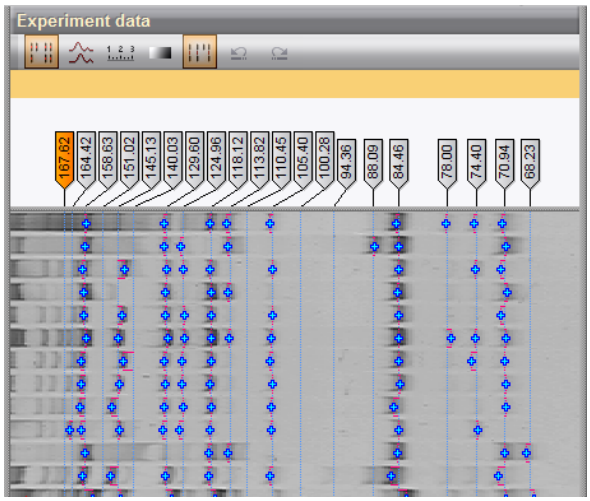


Figure 3.5.3: Band classes.

1.12 To export a tab-delimited text file containing the binary band matching table, select *Composite* > *Export character table* and press <***Yes***>.

1.13 To show the intensity of the bands as colors, choose *Composite* > *Show quantification (colors)* ( ) (see Figure 3.5.4).

1.14 To display the intensities of the bands as numerical values, select *Composite* > *Show quantification (values)*.

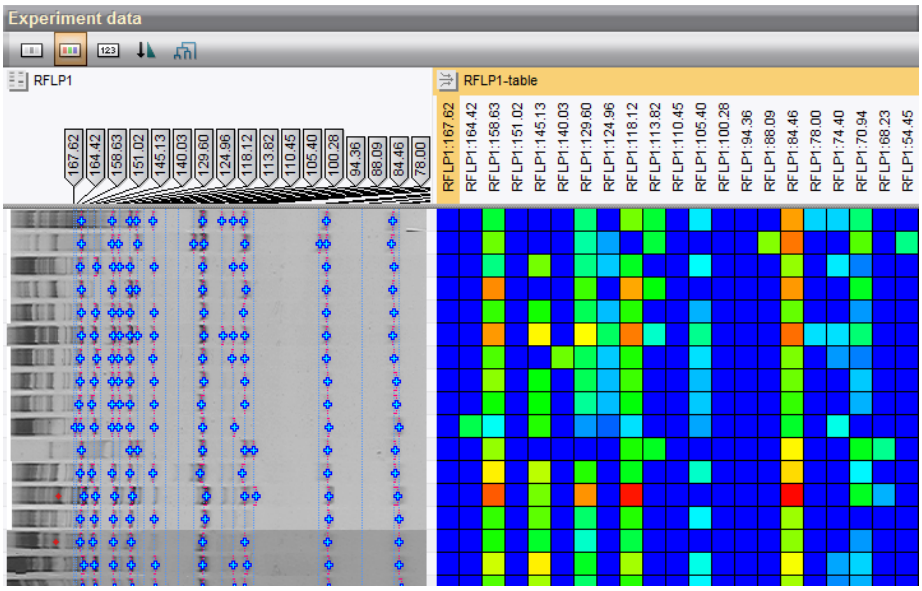


Figure 3.5.4: Left: band classes on the fingerprint type; Right: intensities of the bands shown in color.

3.5.2 Band polymorphism analysis

Band matching tables allow you to identify band classes that discriminate sets of entries. For example, you might want to identify bands that appear only in certain entries. Such discriminating bands can then be analyzed further using other experimental methods.

3.5.2.1 Finding discriminative band classes

Band classes can be sorted based on how well they discriminate a set of entries from the rest. In this example we will find band classes that separate *Vercingetorix* entries from the others.

2.1 Make sure the composite data set **RFLP1-table** is shown and selected in the *Comparison window*.

2.2 Minimize or reduce the *Comparison window* so that the *Information fields window* (at least the menu and toolbar) becomes visible.

2.3 Press **F4** to make sure that no entries are selected.

2.4 Press **F3** and enter **V*** in the **Genus** field of the *Entry search window*. Press <***Search***>.

2.5 To view the selected entries, choose *Edit* > *Arrange entries* > *Bring selected entries to top* in the *Comparison window*.

2.6 Select *Composite* > *Discriminative characters*.

The characters (band classes) are rearranged so that the characters *positive* for the selected entries are to the *left*, and the characters *negative* for the selected entries are to the *right*. Characters in the middle are relatively uninformative with respect to the delineation of Vercingetorix.

3.5.2.2 Sorting entries by band intensity

The band matching table also allows you to sort entries by band intensity. This helps to identify entries for which a band of particular interest is present.

2.7 Select *Composite* > *Show quantification (colors)* ( ) to show the band table as an intensity table.

2.8 Select a band class by clicking on its label in the header. The band class is highlighted.

2.9 Select *Composite* > *Sort by character* or press the button.

The entries are now sorted by increasing band intensity for the selected band class.

3.5.3 Additional practice

3.1 In the **E. coli** database - created in 1.2.1 - create a new composite data set for both fingerprint types **PFGE-XbaI** and **PFGE-AvrII**.

3.2 Perform band matching on entries linked to **PFGE-XbaI** and **PFGE-AvrII** data.

Chapter 3.6

Composite data sets

3.6.1 Introduction

With a composite data set multiple experiments can be combined into a single analysis. Two options are possible for the calculation of the similarity between entries based on a composite data set:

- **Option 1:** The individual similarity matrices are calculated for each experiment type that is present in the composite data set, and a combined matrix is then calculated by averaging the values.
- **Option 2:** All characters from the different experiment types present in the composite data set are merged, and from this set, the similarity matrix is calculated (= ”combined matrix”).

3.6.2 Combining character experiments

In this first example we will combine two different character type experiments.

3.6.2.1 Creating a composite character set

Setting up a composite data set requires just a few steps:

2.1 In the **DemoBase Connected** database, press from the *Experiments panel toolbar* and select *New composite data set*.

2.2 Enter a name (e.g. **Pheno-all**) and press <***OK***>.

The *Composite data set window* is shown for **Pheno-all**.

2.3 Select **PhenoTest** and select *Experiment* > *Use in composite data set*.

2.4 Repeat this for **FAME** (see Figure 3.6.1).

In this composite data set, we will let the software correct for internal weights, since **FAME** contains more characters than the **PhenoTest**.

2.5 Select *Experiment* > *Correct for internal weights*.

2.6 Open the comparison settings for the composite data set by selecting *Experiment* > *Comparison settings*.

2.7 Make sure *Average from experiments* is selected in the first tab and press <***OK***>.

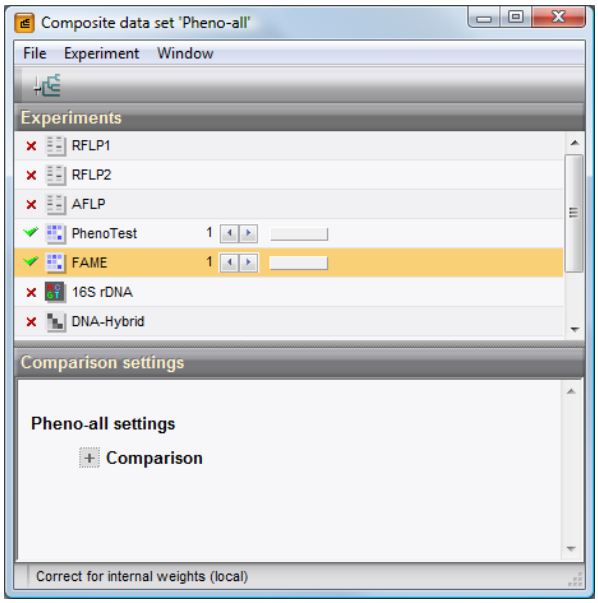


Figure 3.6.1: The *Composite data set window*.

2.8 Select *File* > *Exit* to close the *Composite data set window*.

Pheno-all is listed in the *Experiments panel* of the *BioNumerics main window*.

3.6.2.2 Cluster analysis of a composite character set

Composite data sets allow you to display and arrange the data in novel ways:

2.9 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non-STANDARD entries (see 3.1.3).

2.10 In the *Experiments panel* select **Pheno-all** and press next to **Pheno-all**.

2.11 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

2.12 Make sure *Average from experiments* is selected and press <***Next***>.

2.13 Select *UPGMA* in the next step of the wizard and press <***Next***>.

The resulting dendrogram is shown in the *Dendrogram panel* and is based upon the average matrix of both similarity matrices.

2.14 To show the data as colors, choose *Composite* > *Show quantification (colors)* ( ).

2.15 Select *Composite* > *Calculate clustering of characters* or press the button in the *Experiment data panel*.

2.16 Select *Pearson correlation* and press <***OK***>.

The characters are clustered (see Figure 3.6.2).

2.17 Select a set of entries by pressing the Ctrl-button and clicking on a node in the dendrogram.

2.18 Select *Composite* > *Discriminative characters*.

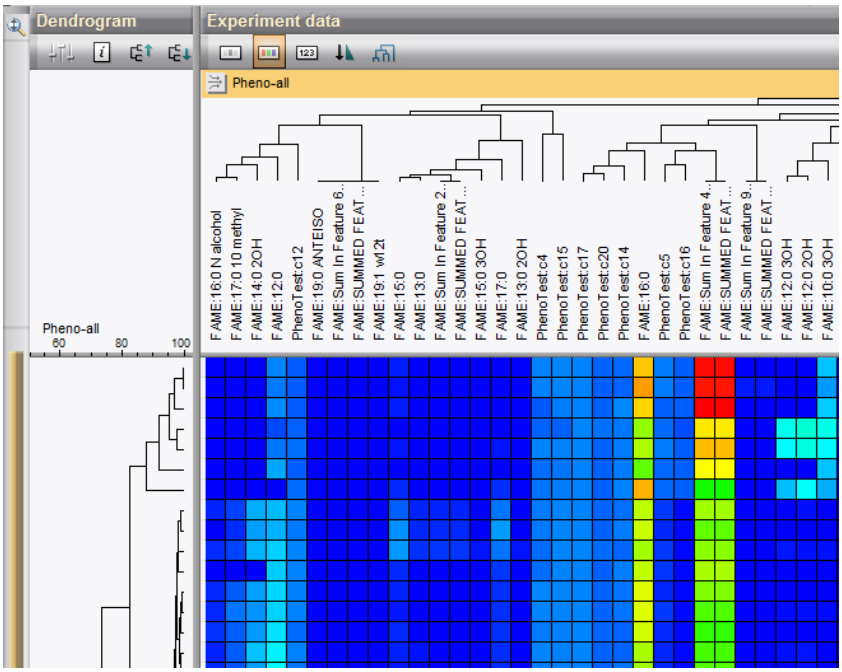


Figure 3.6.2: Transversal clustering: entries (horizontal) and characters (vertical).

The characters are arranged according to how well they discriminate the selected entries from the other entries. Characters at the left are positive indicators, characters at the right are negative indicators, while those in the middle are uninformative.

3.6.3 Combining fingerprint experiments

In this example we will combine two fingerprint type experiments.

3.6.3.1 Creating a composite data set

3.1 In the **DemoBase Connected** database, press from the *Experiments panel toolbar* and select *New composite data set*.

3.2 Enter a name (e.g."RFLP–combined") and press <***OK***>.

The *Composite data set window* is shown for **RFLP-combined** (see Figure 3.6.3).

3.3 Select **RFLP1** and select *Experiment* > *Use in composite data set*.

When an experiment type is selected in the composite data set, it is marked with a green check.

3.4 Repeat this for **RFLP2**.

3.5 Select *File* > *Exit* to close the window.

The new composite data set is listed in the *Experiments panel* of the *BioNumerics main window*.

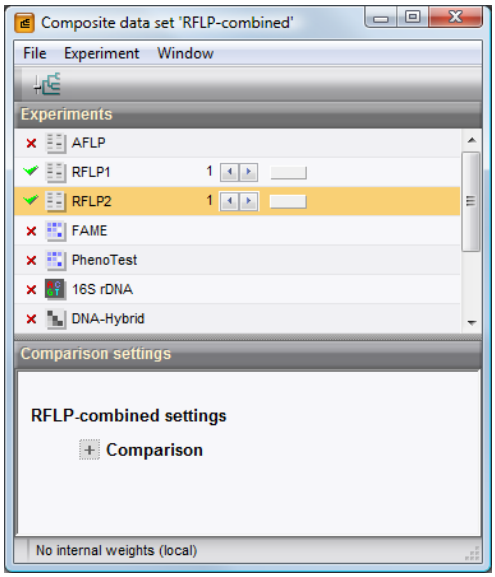


Figure 3.6.3: The *Composite data set window*.

3.6.3.2 Cluster analysis of a composite data set

Option 1: Average the individual matrices

3.6 In the **DemoBase Connected** database, double–click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non–STANDARD entries (see 3.1.3).

3.7 Select **RFLP-combined** in the *Experiments panel* and select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

3.8 Select the option *Average from experiments* and press <***Next***>.

3.9 Leave all settings unaltered in the next step of the wizard and press <***Next***> once more.

With the option *Average from experiments*, the similarity matrices from the individual experiments (**RFLP1** and **RFLP2**) are averaged. The resulting dendrogram is based upon this average matrix.

Option 2: Create combined character matrix

Fingerprints can only be combined to a character matrix if a band matching is performed (see Chapter 3.5).

3.10 In the **DemoBase Connected** database, double-click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non-STANDARD entries (see 3.1.3).

3.11 In the *Comparison window*, select **RFLP1** in the *Experiments panel* and press .

3.12 Select *Fingerprints* > *Perform band matching* or press the button.

3.13 Select *Find classes on all entries* and press <***OK***>.

3.14 Repeat this for **RFLP2**.

3.15 Select **RFLP-combined** in the *Experiments panel*.

The band matching tables of **RFLP1** and **RFLP2** are displayed in the *Experiment data panel*.

3.16 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)* or press and select *Calculate cluster analysis*.

3.17 Select *Dice* and press <***Next***>.

3.18 In the next step, make sure *UPGMA* is selected and press <***Next***> to calculate the cluster analysis.

Both band matching tables are merged to obtain a composite data set. From this composite data set, a similarity matrix is calculated, resulting in a combined dendrogram.

3.19 To show the intensity of the bands as colors, choose *Composite* > *Show quantification (colors)* ( ).

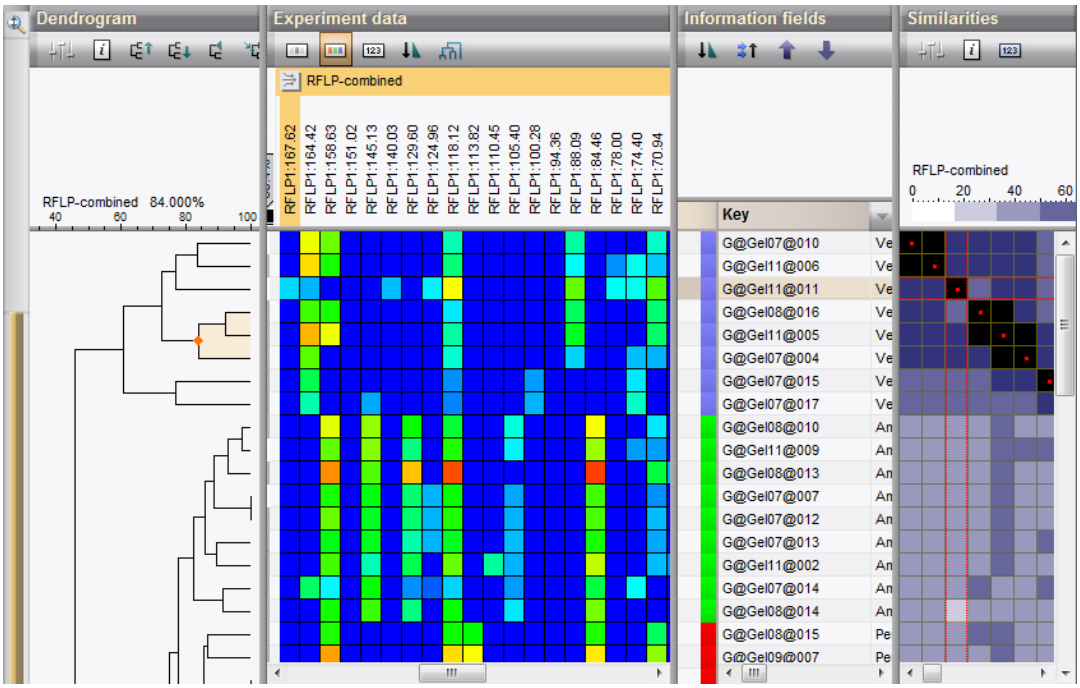


Figure 3.6.4: Cluster analysis of a composite data set containing two fingerprints.

3.20 Save the comparison, and close the window.

3.6.4 Additional practice

Create the following composite data set experiments in the **E. coli** database:

- PFGE-XbaI + PFGE-AvrII
- PFGE-XbaI + PFGE-AvrII + Biolog
- Biolog + Pheno

How would you do cluster analysis on each of these experiments?

Chapter 3.7

Dimensioning techniques

3.7.1 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is an optimized three-dimensional representation of the similarity matrix. The Euclidean distance between two points (entries) reflects the similarity between them as well as possible, while providing a convenient visual interpretation. A similarity matrix must be present before an MDS can be calculated.

3.7.1.1 Calculating an MDS

1.1 In the **DemoBase Connected** database, double–click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non–STANDARD entries (see 3.1.3).

1.2 Select **FAME** in the *Experiments panel* and calculate a dendrogram based on the *Euclidean distance* with *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)*.

1.3 Select *Statistics* > *Multi-dimensional scaling...* ( ).

1.4 Press <***Yes***> to optimize the positions.

The MDS is calculated and the *Coordinate space window* is shown (see Figure 3.7.1). The *Coordinate space window* shows the entries as dots in a cubic coordinate system.

1.5 To zoom in and zoom out on the image, press the **PageDown** and **PageUp**–keys, respectively. Alternatively, the zoom slider can be used.

1.6 The image can be rotated in real time by clicking on the image and dragging the mouse in the desired direction.

By default, the entries appear in the colors as defined for the groups in the *Comparison window*.

1.7 If no groups are defined in the *Comparison window*, right-click on the database field name **Genus** in the *Information fields panel*, and select *Create groups from database field*. Select the order in which groups are created (i.e. by size, alphabetically, or by position in the comparison) and press <***OK***> to create the groups.

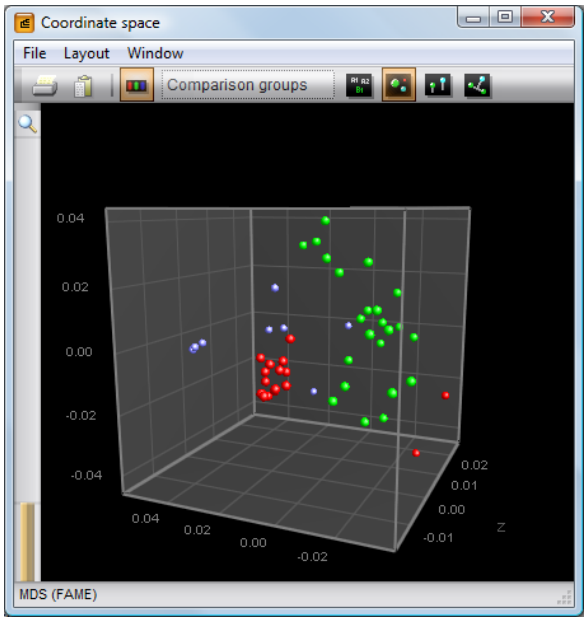


Figure 3.7.1: The *Coordinate space window*.

3.7.1.2 Changing the coordinate space layout

MDS is a visualization tool, and there are several ways to modify its appearance.

1.8 With *Layout* > *Show keys* ( ), you can display the database keys of the entries.

1.9 In the *Comparison window*, select *Layout* > *Use group numbers as keys*.

The entries in the *Comparison* and in the *Coordinate space window* are now labeled with a group-specific letter and an entry-specific number.

1.10 Alternatively, you can select a field in the *Comparison window*, for example the **Strain number** field, and select *Layout* > *Use field as key*.

1.11 A list of entry labels as used in the MDS and corresponding database fields can be exported by selecting *File* > *Export* > *Export database fields* in the *Comparison window*.

1.12 With *Layout* > *Show group colors* ( ), you can toggle between the color representation and the non–color representation, in which the entry groups are represented (and printed) as symbols instead of colored dots.

1.13 With *Layout* > *Show construction lines* ( ), the entries are displayed on vertical lines starting from the bottom of the cube. This may facilitate the three-dimensional perception.

1.14 With *Layout* > *Show rendered image* ( ), you can toggle between the realistic three–dimensional perspective with entries represented by spheres, and a simple mode where entries are represented as dots.

1.15 Select *Layout* > *Show dendrogram* ( ) to show the relatedness among entries as defined by the dendrogram.

1.16 Select *File* > *Print image...* ( ) to print the image. The image will print in color if the colors are shown on the screen.

3.7.2 Principal components analysis (PCA)

Principal components analysis (PCA) is another way to visualize relationships among entries. Instead of using the similarity matrix to measure relatedness, PCA uses the data set itself. Mathematically, entries can be plotted in N-dimensional space, where each dimension corresponds to one of N characters, and each entry's position corresponds to its N character values. If there are more than three characters, then this plot becomes impossible to visualize. PCA reorients the plot to maximize the variation among entries along the first two or three dimensions, which can then be displayed. These are the principal components.

3.7.2.1 Calculating a PCA

Since PCA operates on the data set, rather than on the similarity matrix, no preliminary cluster analysis is necessary. However, it is essential that the experiment contains *well–defined characters*, whether they are band classes or characters in a character set. PCA can also be done on aligned sequences.

2.1 In the **DemoBase Connected** database, double–click on the comparison **All** in the *Comparisons panel* of the *BioNumerics main window* to open the *Comparison window*.

The comparison **All** contains all non–STANDARD entries (see 3.1.3).

2.2 If no groups are defined in the *Comparison window*, right–click on the database field name **Genus** in the *Information fields panel*, and select *Create groups from database field*. Select the order in which groups are created (i.e. by size, alphabetically, or by position in the comparison) and press <***OK***> to create the groups.

2.3 Select **FAME** in the *Experiments panel* and select *Statistics* > *Principal Components Analysis* or press the button.

2.4 Check *Subtract average* in the *Characters panel*, and leave the other options unchecked.

2.5 In the *Component type panel*, select *Principal components*, and press <***OK***>.

The *Principal Components Analysis window* pops up. The *Principal Components Analysis window* is divided into three panels (see Figure 3.7.2).

- The *Entry coordinates panel* shows the entries plotted in two dimensions corresponding to the first two principal components (X and Y).
- The *Character coordinates panel* shows the characters plotted in the same two dimensions.
- The *Components panel* lists the principal components in the order of their contribution to overall variance. The components used as X, Y and Z axes are also indicated.

A character that appears near the edge of the plot is a *strong* discriminator, while a character near the center is a *weak* discriminator. Furthermore, a character that appears near the position of an entry is an *indicator* for that entry.

3.7.2.2 Changing the PCA layout

Principal components analysis in BioNumerics is essentially a visualization tool, and its appearance can be modified in several ways.

2.6 If you have assigned groups, you can switch from colors to symbols to indicate the groups by pressing .

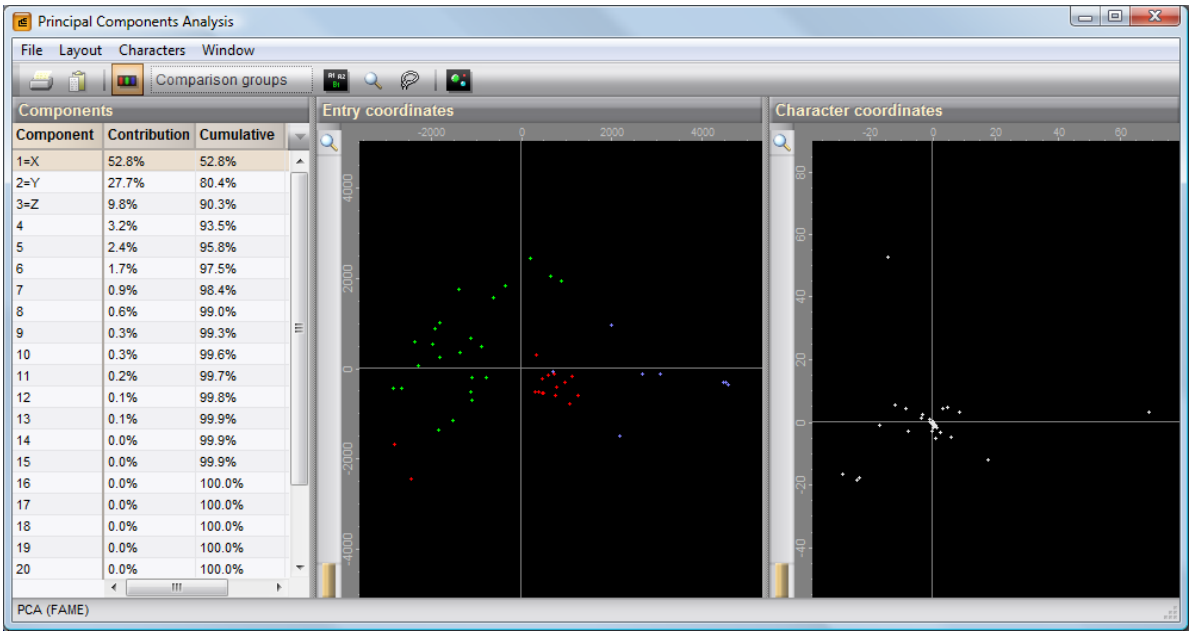


Figure 3.7.2: The *Principal Components Analysis window*

2.7 To show the keys or unique labels for the entries, press .

2.8 To view another component in the plot, select that component in the components bar and then select *Layout* > *Use component as X axis* or *Use component as Y axis*.

2.9 To zoom in on any part of the PCA plot, press . Then drag the mouse pointer to create a rectangle. The area within the rectangle will be expanded to cover the whole panel.

2.10 In order to restore the original size of the image, left-click within the panel. Press to disable the zoom-mode.

2.11 Move the mouse pointer over the characters in the right panel to see their names.

2.12 Entries can be selected in a *PCA window* by holding the **Shift**–key down and drawing a rectangle around the entries with the left mouse button. Selected entries are circled in blue.

2.13 Select *File* > *Print image (entries)* to print the entry plot, or select *File* > *Print image (characters)* to print the character plot.

2.14 Select *File* > *Copy image to clipboard (entries)* to copy the entry plot to the clipboard, or select *File* > *Copy image to clipboard (characters)* to copy the character plot to the clipboard.

2.15 To create a three–dimensional plot from the PCA, press .

2.16 Select *File* > *Exit* to close the *Coordinate space* and the *Principal Components Analysis window*.

Part 4

Identification

Chapter 4.1

Identification of unknown entries

4.1.1 Identification of unknown entries in a comparison

Simple identifications can be done within the *Comparison window*. In the **DemoBase Connected** database, we will identify unknown species in the genus **Ambiorix**.

1.1 Open the **DemoBase Connected** database.

1.2 Right–click on the 'Species' field name in the main database, and select *Arrange entries by field*.

1.3 Right–click on the 'Genus' field name in the main database, and select *Arrange entries by field*.

Now, the entries are arranged first by 'Genus', then by 'Species'.

1.4 Select all **Ambiorix** entries (click on the top **Ambiorix** entry then press the **Shift**–button while clicking on the bottom **Ambiorix** entry).

1.5 Select *Comparison* > *Create new comparison* (**Alt+C**) or press the button from the *Comparisons panel* toolbar.

1.6 In the *Comparison window*, resize the *Similarities panel* on the right to make space for the similarity values, if necessary.

1.7 In the *Experiments panel*, select **RFLP1** as the experiment to identify the unknown entries.

1.8 Select an **Ambiorix sp.** in the *Information fields panel*.

1.9 In the menu of the *Comparison window*, choose *Edit* > *Arrange entries* > *Arrange entries by similarity* ( ).

The selected entry appears at the top followed by all other entries in the comparison. They are arranged by decreasing similarity to the selected entry. The similarity values are shown in the *Similarities panel* (see Figure 4.1.1).

4.1.2 Identification of unknown entries in a library

Library identification in BioNumerics is essentially a series of automated comparisons. Libraries can be organized by taxonomic units, such as species, or by any other category of interest. Unlike the comparison method described above, library identification is based upon comparisons to each unit as a whole, rather than to individual entries.

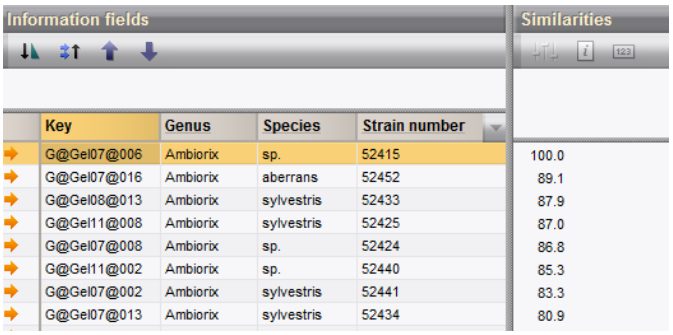


Figure 4.1.1: Entries are arranged by decreasing similarity.

4.1.2.1 Creating a library

A library is made up of units, each of which is analogous to a comparison. In the **DemoBase Connected** we will create a library based on the different species.

2.1 In the **DemoBase Connected** database window, press the *Libraries tab* to show the panel (see Figure 4.1.2).

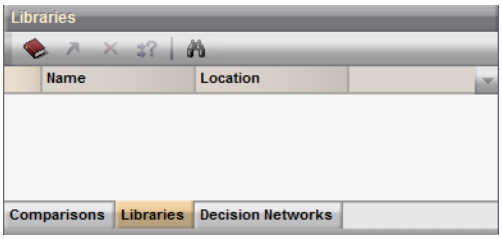


Figure 4.1.2: The *Libraries panel*.

2.2 Select *Identification* > *Create new library...* ( ).

2.3 Enter a name for the library, such as **DemoLib**.

The *Library window* pops up (see Figure 4.1.3).

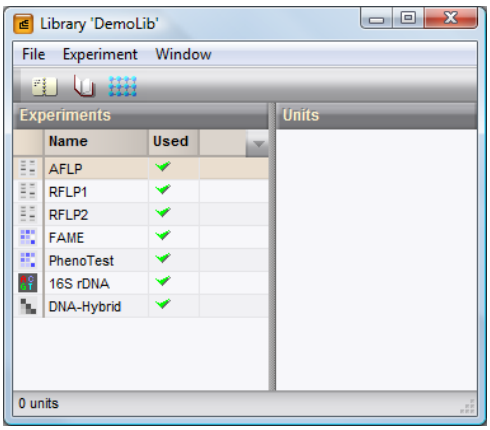


Figure 4.1.3: The *Library window* of a new library.

2.4 Select *File* > *Add new library unit...* ( ).

2.5 Enter a name of one of the species in the database, for example ”Ambiorix sylvestris”.

The library unit now shows up in the *Units panel* of the *Library window*.

2.6 Double-click on the unit to open it.

The *Unit window* which appears, is very similar to the *Comparison window*.

2.7 In the *BioNumerics main window*, select all **Ambiorix sylvestris** entries. Use the **Shift**–key to select a range of entries.

2.8 Select *Edit* > *Copy selection* (, **Ctrl+C**) to copy the selected entries to the clipboard.

2.9 In the *Unit window*, select *Edit* > *Paste selection* (, **Ctrl+V**) to paste the entries into the window.

2.10 Save the library unit with *File* > *Save* (, **Ctrl+S**) and close the unit.

2.11 Repeat these steps to create a library unit for each of the other species: Ambiorix aberrans, Perdrix pseudoarchaeus, Vercingetorix aquaticus, Vercingetorix nemorosum, and Vercingetorix palustris.

The *Library window* should now contain six units, each having their own set of entries (see Figure 4.1.4).

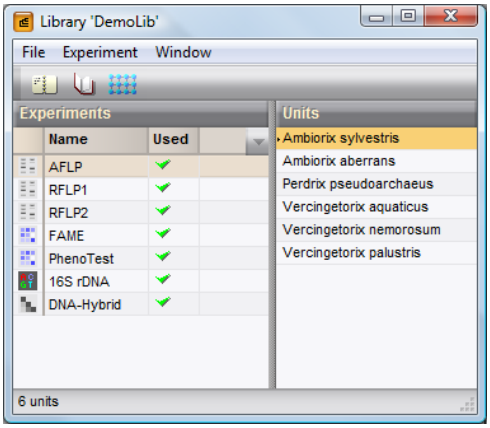


Figure 4.1.4: The *Library window* with six units.

4.1.2.2 Identifying unknown entries with a library

Now that we have a library with units based on the species, we can identify entries.

2.12 Select a list of entries, for example all unnamed species and a few entries of the other species.

2.13 Select the **DemoLib** in the *Libraries panel* to specify that library for identification.

2.14 Select *Identification* > *Identify selected entries*.

2.15 In the *Identification dialog box*, select *Mean similarity* and check *Calculate normalized distances*.

2.16 Press <***OK***> to start the calculations.

The *Identification window* appears, showing the progress of the calculations in the progress bar in the bottom of the window. Once the calculations are done, the window is divided in three panels (see Figure 4.1.5):

- The unknowns are listed in the *Unknowns panel*.

- For each unknown, the best matching library unit is listed in the *Matches panel*.
- The *Details panel* contains the scores and normalized distances for the selected unknown.

2.17 Press the button to shown the second best match for each unknown.

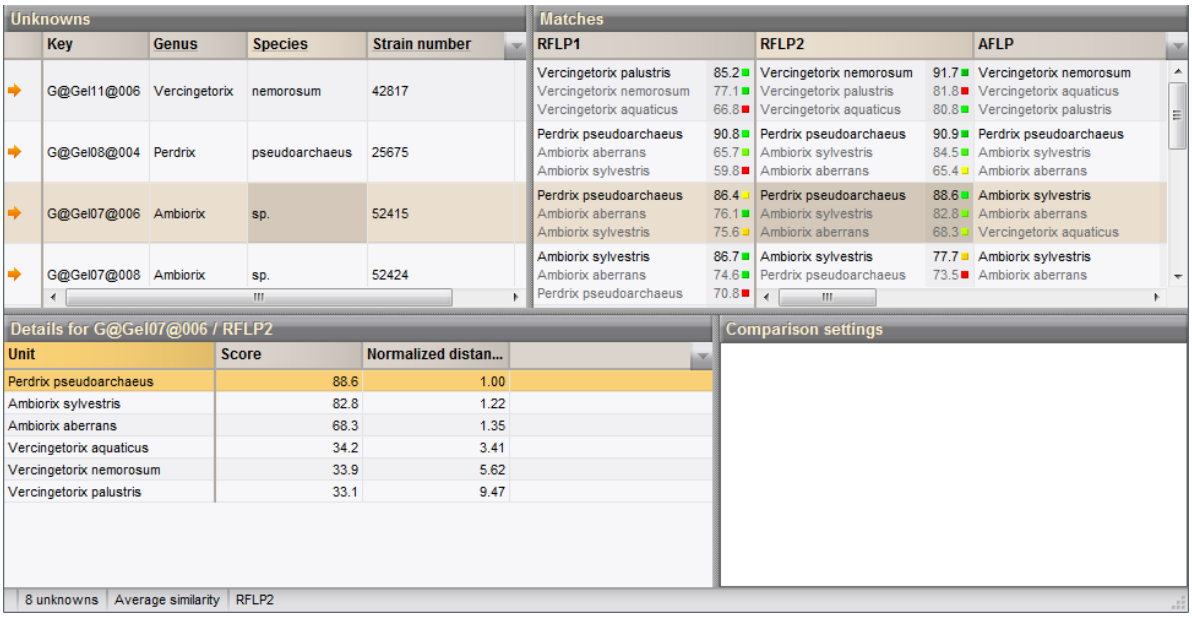


Figure 4.1.5: The *Identification window*.

The normalized distances appear as colored dots next to the similarity scores, ranging from red (poor) to green (excellent). They indicate how well the unknown fits into the unit, relative to other members of the unit.

Chapter 4.2

Decision networks

4.2.1 Introduction

Decision networks should be seen as a construction kit that allows you to build your own automated decision or action work flows, with practically endless possibilities. They can be used to make decisions, predict features, perform queries, fill in fields, create graphs and plots, and much more.

4.2.2 Creating a new decision network

In the default configuration of the *BioNumerics main window*, the *Decision Networks panel* is seen as a tab behind the *Comparisons panel*.

2.1 In the **DemoBase Connected**, click on the tab to bring the *Decision Networks panel* to the top.

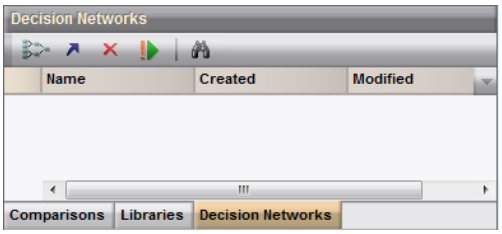


Figure 4.2.1: The *Decision Networks panel*.

2.2 Select *Identification* > *Decision networks* > *Create new* or press the button in the toolbar of the *Decision Networks panel* to create a new empty decision network.

2.3 Enter a name in the dialog box that pops up, for example, **My DN**.

The new decision network is now listed in the *Decision Networks panel*. When a decision network is opened, it contains by default the current selection of entries. Therefore, it is practical to make a selection of entries you want to use in the decision network before opening it.

2.4 As an example, select all entries except the ones marked as ”STANDARD” in the **Genus** field.

2.5 Open the decision network by pressing the button or by double-clicking on its name.

The *Decision network window* contains four panels:

- The main *Network panel* displays the network scheme.

- The *Operators panel* lists a tree of all operators that are available to construct the decision network (the building blocks).
- In the *Node properties panel*, the properties and data of the current selected node is given.
- The *Entry data panel* lists the entries currently used in the network and their selection status. In the right hand sub-panel, the output(s) from the network are listed for the entries (currently empty).

4.2.3 Building a decision network

As an example, we will create a simple decision network that discriminates between the three genera in **DemoBase Connected**, based upon the 16S rDNA sequences.

3.1 In the newly created decision network, open the *Data sources* group in the *Operators panel* by clicking on its icon.

3.2 Double-click on *Sequence*, which opens a *New operator dialog box* (Figure 4.2.2).

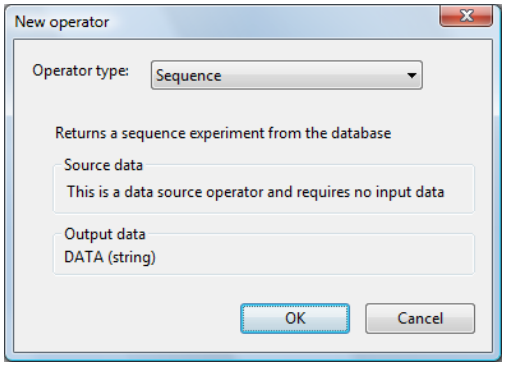


Figure 4.2.2: The *New operator dialog box* for a Sequence operator.

The dialog box describes the operator and mentions the source data needed and the output data delivered to the network.

3.3 Press <***OK***> to edit the node properties for the sequence input node.

3.4 Enter e.g. ”16S” as a *Name*, select **16S rDNA** (the only sequence type available in **DemoBase Connected**) as *Sequence type* and press <***OK***>.

The network now contains one node, i.e. ”16S”.

3.5 If you click on an entry in the *Entry data panel*, the node and the *Node properties panel* are updated with the sequence data of the highlighted entry.

3.6 Select the node ”16S” in the network (a selected node is bordered by a red line).

3.7 Open the *Sequence operators* group in the *Operators panel* and double-click on *Find subsequence*. Press <***OK***>.

3.8 Enter **Ambiorix** as *Name*, and enter ”GGGTGTAG” as *Match sequence*. Press <***OK***> to confirm the node properties.

The network is now ready to produce a first result.

3.9 In the *Decision network window*, press the button to calculate the network.

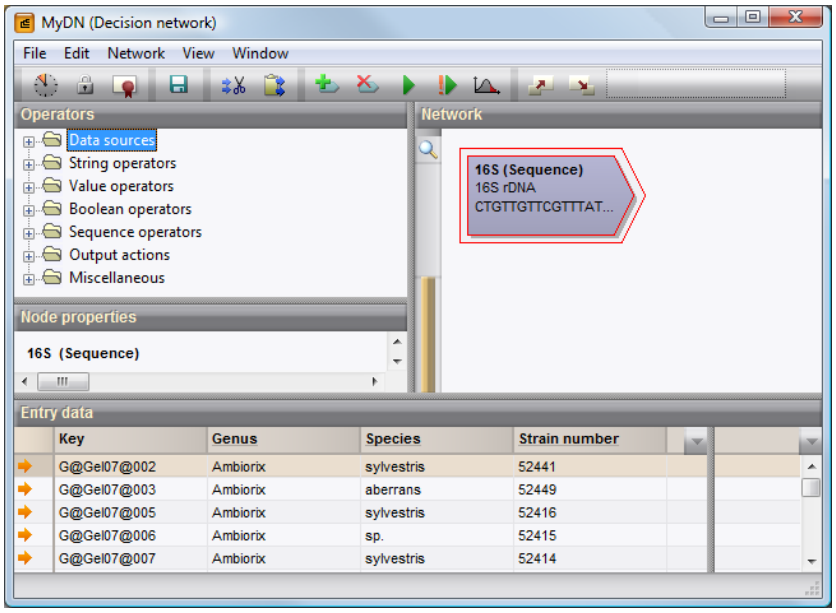


Figure 4.2.3: The *Decision network window*.

The percentage of true and false entries in the *Entry data panel* is indicated as a green and red bar, respectively.

3.10 Continue to build the network by selecting the data node "16S" again, click on the *Sequence operators* group in the *Operators panel* and double-click on *Find subsequence*. Press <***OK***>.

3.11 Enter **Vercingetorix** as *Name*, and enter "CGATCTCACG" as *Match sequence*. Press <***OK***> to confirm the node properties.

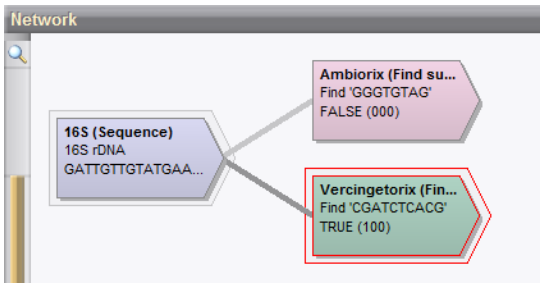


Figure 4.2.4: Two boolean output nodes.

3.12 Press the button to calculate the network.

3.13 Select both output nodes by clicking the first and then, while holding the Ctrl-key, clicking the second. Both nodes are now bordered in red.

3.14 Combine the two nodes with the *OR* operator from the *Boolean operators group*. Press <***OK***> twice.

3.15 Select the *NOT* operator from the *Boolean operators* group. Press <***OK***>, give the node the name **Perdrix** and press <***OK***> once more.

3.16 Press the button to calculate the network.

The colors for the nodes depend on the entry that is selected in the *Entry data panel*.

3.17 Select the following three nodes: **Ambiorix**, **Vercingetorix**, and **Perdrix**. Use the Ctrl-button to select the nodes.

3.18 Create a new node using the Boolean operator *Categorical combiner*. The output for this node is a string, containing the category (Ambiorix, Vercingetorix, Perdrix) that is true.

3.19 Enter **Genus** as *Name* and check *Use as output*. Press <***OK***>.

3.20 Press the button to calculate the network.

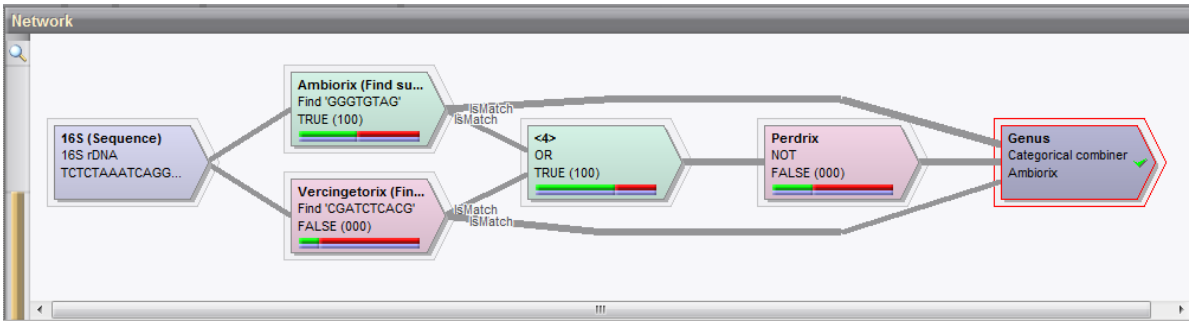


Figure 4.2.5: Decision network that decides between three groups.

The *Entry data panel* contains a new output column **Genus** showing the genus name.

3.21 Select the **Genus** node in the *Network panel* and select the button.

A bar graph appears, showing the occurrences of the three genera.

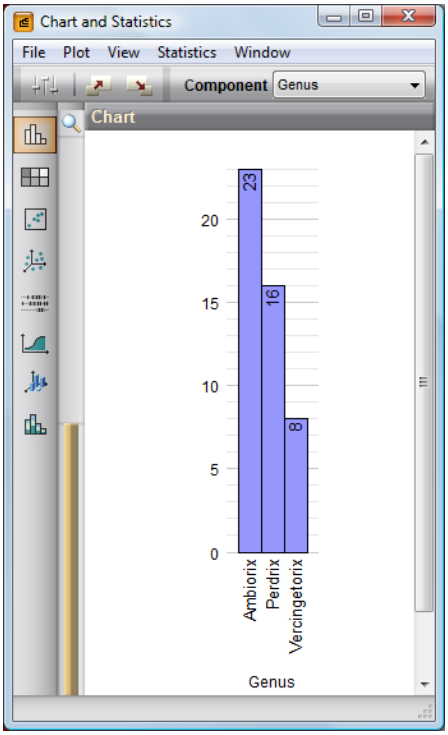


Figure 4.2.6: Bar graph.

3.22 Close all windows and save the project.

www.applied-maths.com

Keistraat 120, B-9830 Sint-Martens-Latem, Belgium
Phone +32 9 2222100, Fax +32 9 2222102

13809 Research Blvd., Suite 645, Austin, Texas 78750, USA
Phone +1 512-482-9700, Fax +1 512-482-9708

APPLIED MATHS