BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

HDA plugin

PLUGINS

VERSION 7.6

www.applied-maths.com

Contents

NOTES

SUPPORT BY APPLIED MATHS

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

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

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

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

LIMITATIONS ON USE

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

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

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

Chapter 1

Starting and setting up BioNumerics

1.1 Introduction

This guide is designed as a tutorial for the *Hetero-duplex Analysis (HDA) plugin* of BioNumerics. The features of the plugin will be illustrated using three sets of 96 hetero-duplex PCR profiles run on a 3100 Genetic Analyzer (Applied Biosystems, Inc., Foster City, CA), obtained by courtesy of Dr. Stephen Edwards, the Institute of Cancer Research, Surrey, UK. The files can be downloaded from the download page on the Applied Maths website (`http://www.applied-maths.com/download/sample-data`, click on "CSCE sample data files").

The minimal configuration for the installation of the *HDA plugin* includes the Fingerprint data module, Character data module, and Dimensioning and Matrix Mining module.

1.2 Startup program

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

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

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

1.3 Creating a new database

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

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

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

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

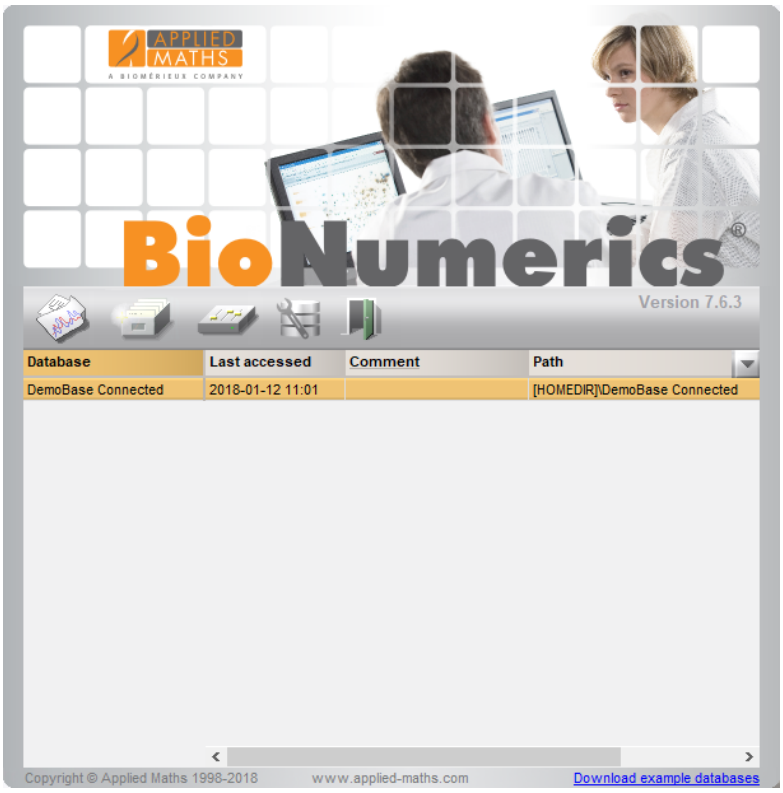

Figure 1.1: The *BioNumerics Startup* window.

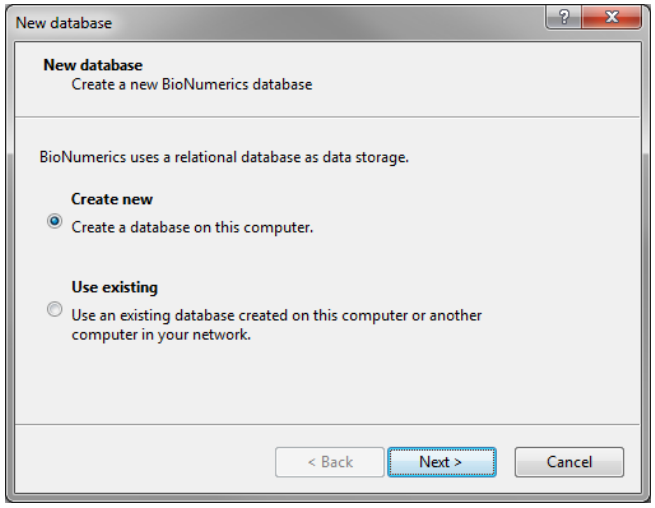

Figure 1.2: The *New database* wizard page.

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

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

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

The *Plugins* dialog box appears.

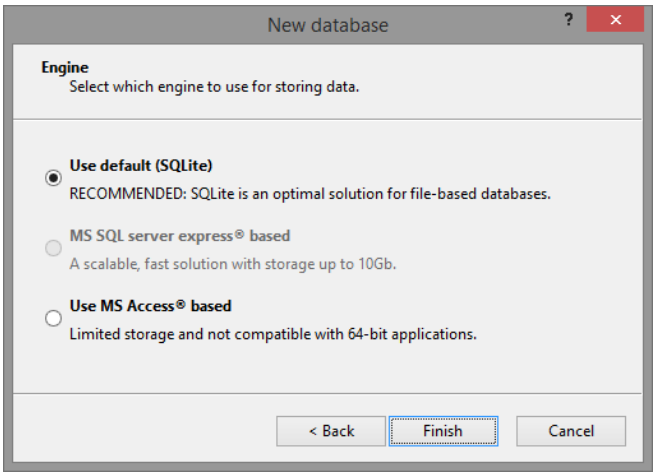

Figure 1.3: The *Database engine* wizard page.

1.4 Installing the HDA Plugin

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

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

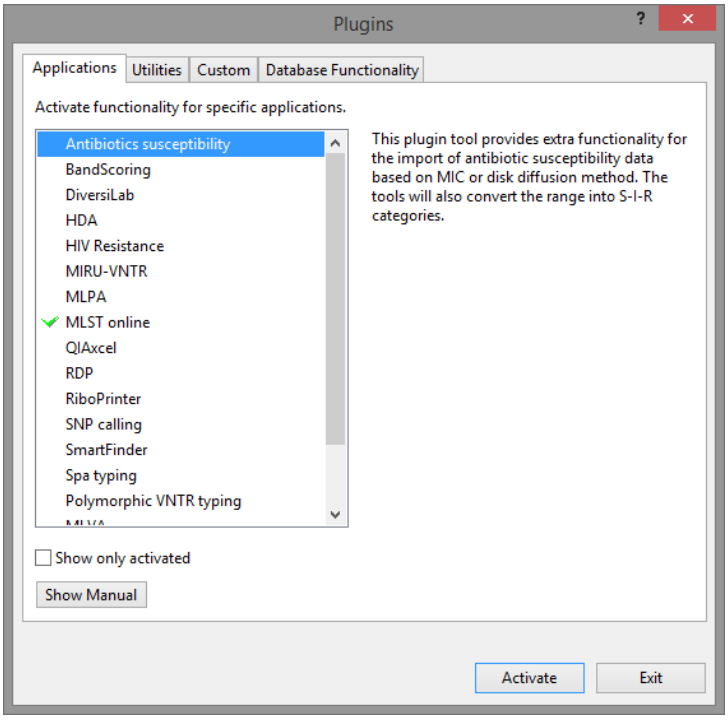

Figure 1.4: The *Plugins* dialog box.

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

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

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

button.

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

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

The installation of the *HDA plugin* requires administrator privileges.

The *HDA plugin* can only be installed and activated with a valid *license number*, which needs to be purchased from Applied Maths.

The *License string dialog box* prompts for a license string that is compatible with the ***Software Serial Number*** listed in the dialog box (see Figure 1.5).

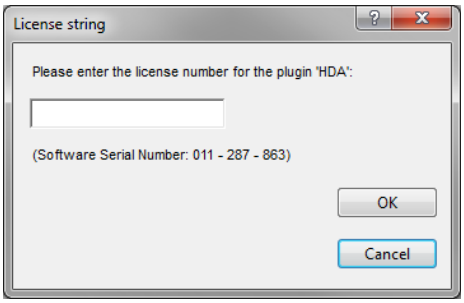

Figure 1.5: The *License string dialog box*.

4.2 Enter the 6 digits License Number and press <***OK***>.

4.3 Press <***OK***> once more to confirm the installation of the *HDA plugin*.

At this point, the *HDA plugin* is installed, and a number of tables is automatically generated in the current database. If any error messages occur, you may not have sufficient privileges to create new tables/columns in the database. Please contact your system administrator in that event.

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

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

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

The *HDA plugin* installs itself in a menu of the BioNumerics software (see Figure 1.6).

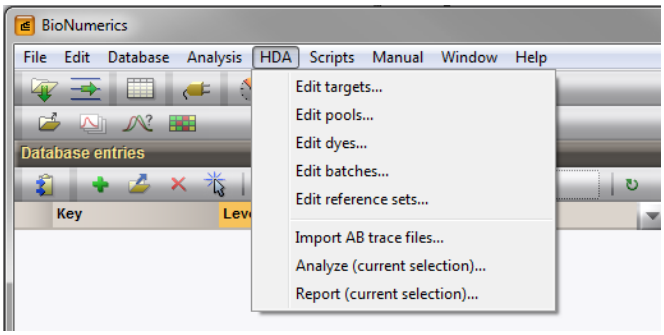

Figure 1.6: HDA menu items in the *Main* window.

Chapter 2

The HDA database

2.1 Elementary database structure

The core object in a BioNumerics database is the *entry*. A database entry is a single object for which experimental information is available. In most cases, the entry will be an organism, a bacterial strain, a sample (tissue, blood, water, ...), but it can also be a complex mixture of organic material, for example in an ecological study.

Given a list of *entries*, a BioNumerics database consists of two parts: on the one hand, there is a number of *information components* that describe the entries, and on the other hand there is *experimental data* that has been obtained for the entries. A scheme of the essence of a BioNumerics database is given in Figure 2.1. The link between the entries, their information fields and their experiment data is established through a *key* field. If the key field is the same for an entry and an experiment, the software knows that the experiment belongs to the entry.

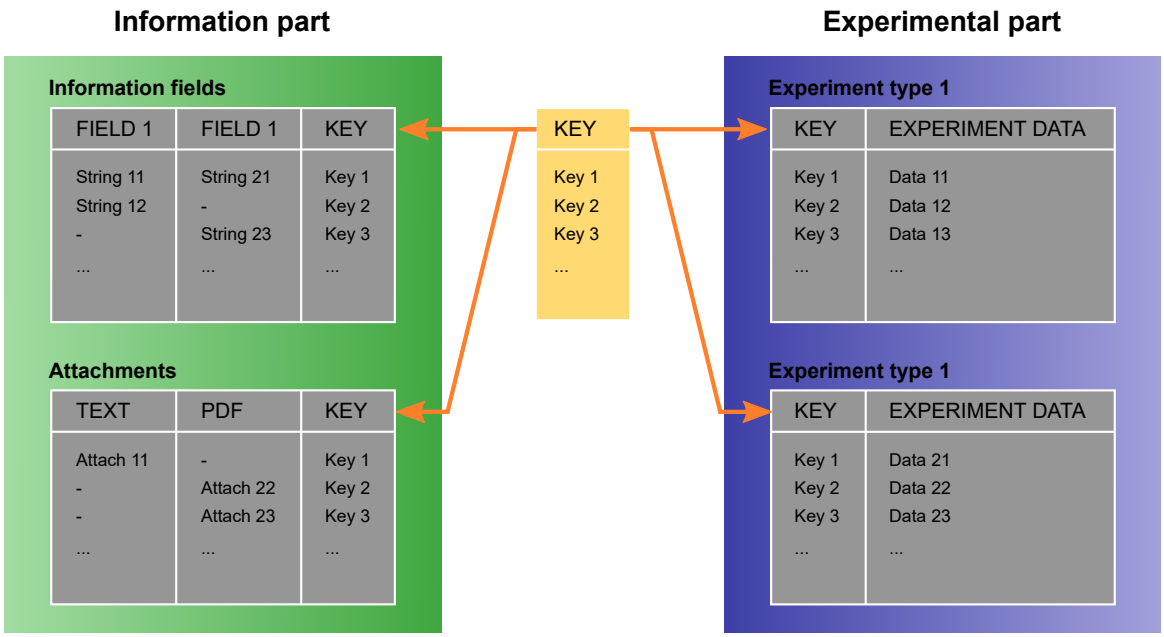

Figure 2.1: Simplified structure of a BioNumerics database.

2.2 Types of experiment data

BioNumerics recognizes different classes of data, which are called *Experiment types*. The HDA application uses two types of experimental data:

- The **Fingerprint type** for storing and processing the densitometric profiles extracted from the .fsa files.
- The **Character type**, to store the results obtained after analysis and comparison of unknown peaks with reference peaks.

2.3 The HDA experimental setup

In a typical HDA experimental setup, a number of PCR *targets* (i.e. target genes on the genome) are amplified to form one *pool* of fragments (sometimes called panel). This can happen through multiplex PCR or the amplification products can be mixed after PCR. The number of PCR products (*targets*) one can pool together depends on (1) the number of color dyes used: if 4 color dyes are used, 4 differently labeled fragments can be pooled, and (2) the possibility to combine fragments with significantly different lengths. Figure 2.2 shows a typical setup where 4 PCR products (*targets*) are mixed in one pool, using 4 color dyes. Since the total number of genes (targets) per sample to be tested is 8, two pools are generated, each containing 4 target PCR products.

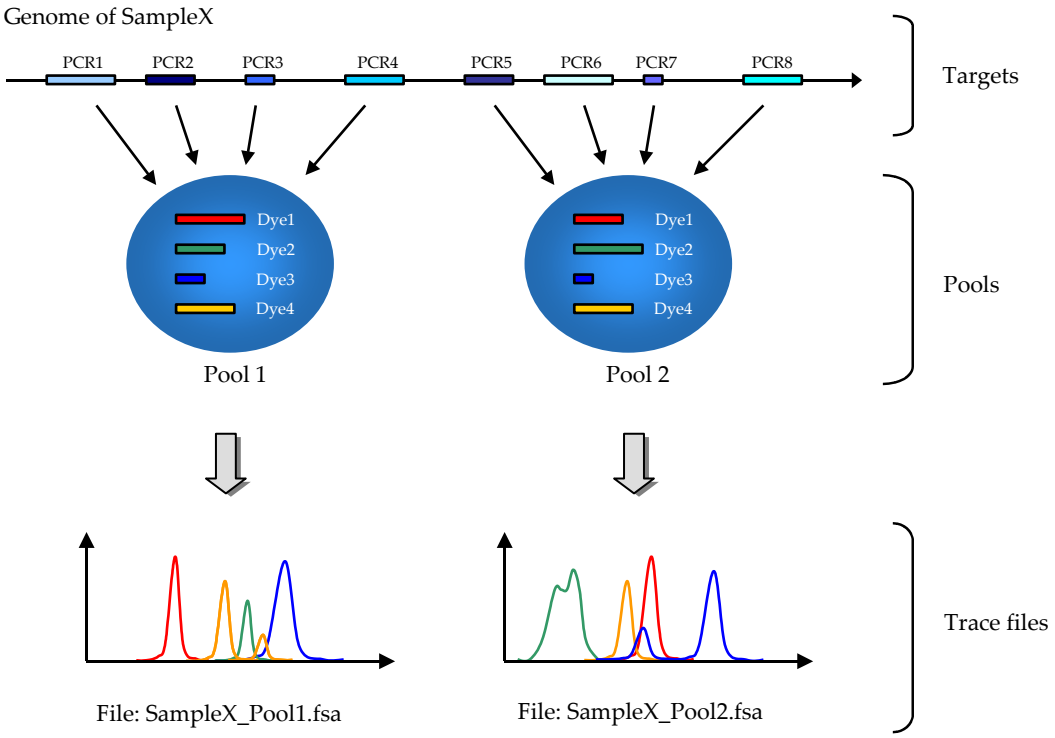

Figure 2.2: Schematic overview illustrating the relation between PCR targets, pools and trace files in a typical HDA experimental setup. In this example, only one target gene is used per dye in the same pool.

An alternative set-up is given in Figure 2.3, where targets (PCR products) have been carefully chosen so that each time a short and a long fragment can be labeled with the same dye and mixed in the same pool. This setup saves on reagents and runs but has the disadvantage that secondary peaks caused by mutant heteroduplexes are more difficult to interpret: if a secondary peak has a length that falls in between the lengths of

the two primary peaks in the same dye, it may be incorrectly interpreted, i.e. assigned to the wrong target. The software has a built-in warning system for such cases. The sample dataset used in this manual is an example of the set-up illustrated in Figure 2.3.

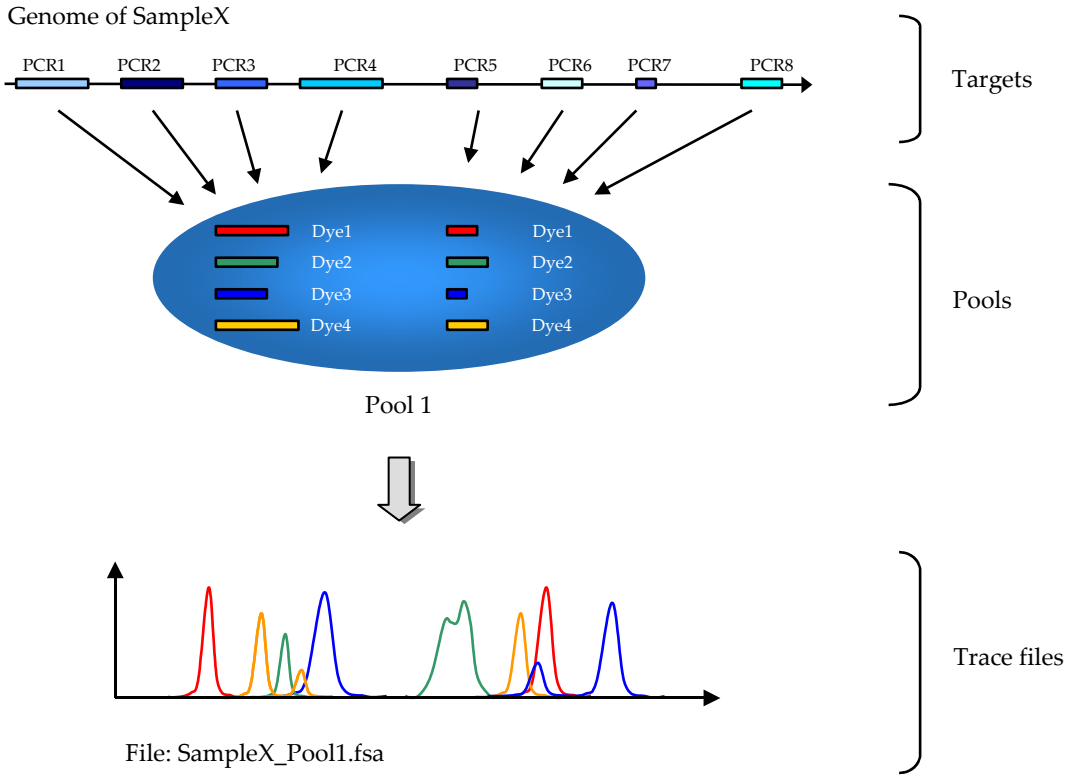

Figure 2.3: Schematic overview illustrating the relation between PCR targets, pools and trace files in an alternative HDA experimental setup. In this example, two compatible targets (i.e. with different lengths) are labeled with the same dye and mixed in one pool.

Trace files (.fsa) are imported in BioNumerics as a *batch*. For example, one complete run of 96 trace files can be imported as one *batch*. Although it is not necessary to pair batches to runs on a sequencer (a batch can contain between 1 and 300 trace files from any origin) it is certainly convenient and logical if batches are created from runs. Figure 2.4 shows a conversion of trace files (.fsa) into fingerprint files. In this example, pools from the same sample are batched. The user enters a file name for the batch of trace files. Technically, the software splits the batched trace files into separate fingerprint files per dye. Therefore, each batch results in one fingerprint file per dye. A fingerprint file gets a file name entered by the user plus an extension referring to the dye number: "_1", "_2", "_3", "_4", "_5",...

2.4 The HDA database structure

The database structure generated by the *HDA plugin* of BioNumerics is specifically designed to store all the information on the peaks (*signals*), *targets*, and the way the targets are organized in *pools*.

The HDA install script creates a number of additional tables in the database, which are required by this plugin. An overview of the table structure specifically related to the *HDA plugin* is shown in Figure 2.5. Note that this table overview is not complete, a number of additional columns and tables are not important to understand the database strategy.

The core component of the HDA database part is the *signal*, i.e. an individual peak resulting from a PCR target. The central table in the database is HDASIGNALS, which contains a record for each *signal*. Other important tables are HDATARGETS, defining the *targets* (PCR products) available in the database, and

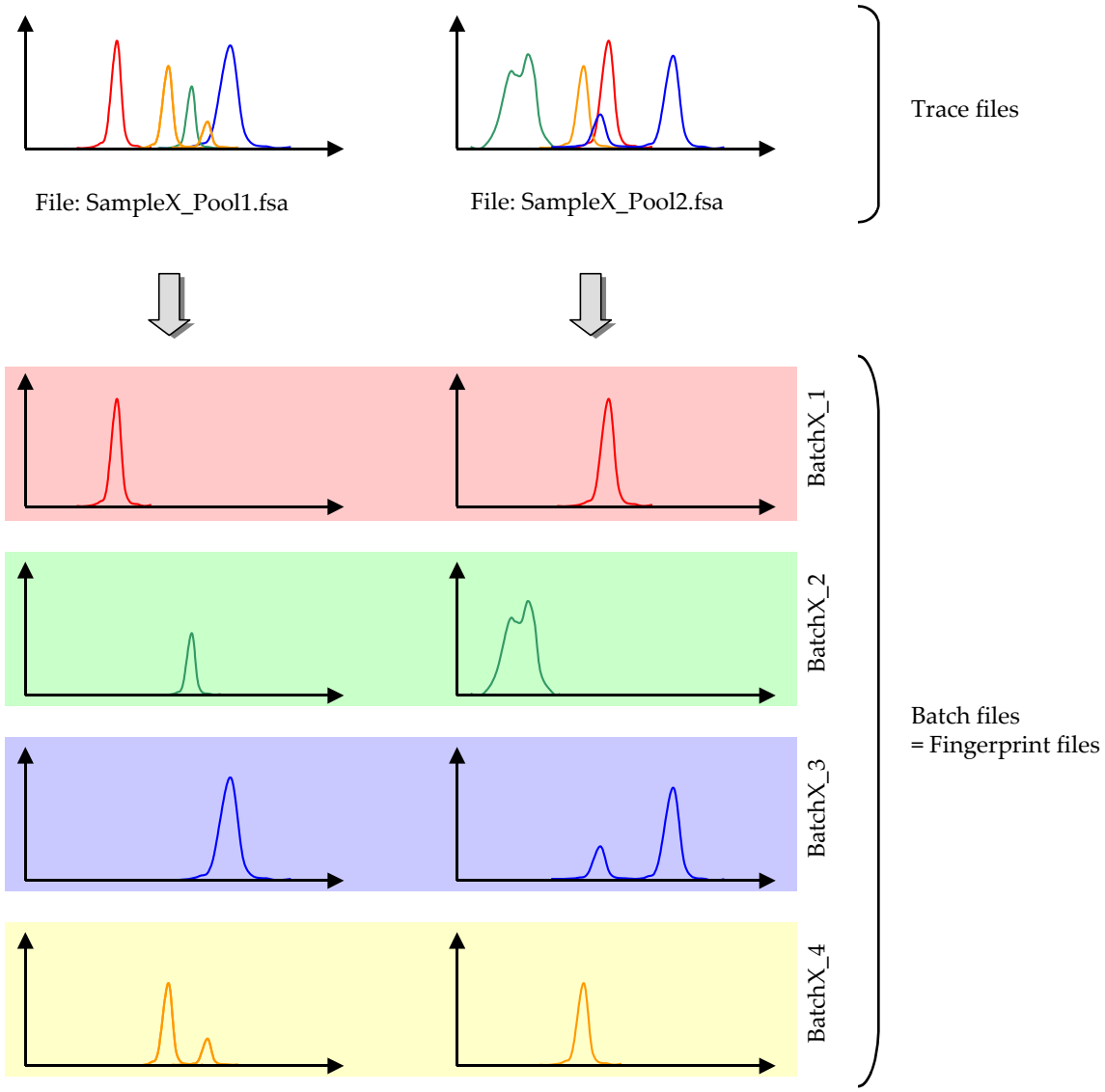

Figure 2.4: Conversion of batches of trace files (.fsa) into fingerprint files.

HDAPOOLS, defining the *pools* (panels) present in the database. ENTRYTABLE contains the *samples* studied, i.e. the DNA samples from which PCR products have been generated.

In the core table HDASIGNALS, each *signal* has a unique ID, SIGNALID. It is characterized by a target-identifier (TARGETID), that corresponds to the target-identifier in table HDATARGETS. This field determines which *target* the *signal* represents. The curve-identifier, CURVEID provides an index number of the profile in a batch imported in BioNumerics, corresponding to the file-index identifier (FILEIDX) in the table of *Fingerprints* (FPRINT). As *fingerprints* are batched in files, the FPRINT table is linked to another table, FPRINTFILES which describes the *fingerprint files*. A *signal* also has a sample-identifier SAMPLEID which tells the software to which sample it belongs, described in the table ENTRYTABLE. The latter table also provides additional information fields for the samples.

Another table, HDAPOOLTARGETS, describes for each target what pool(s) it is used in, and what dye it is labeled with.

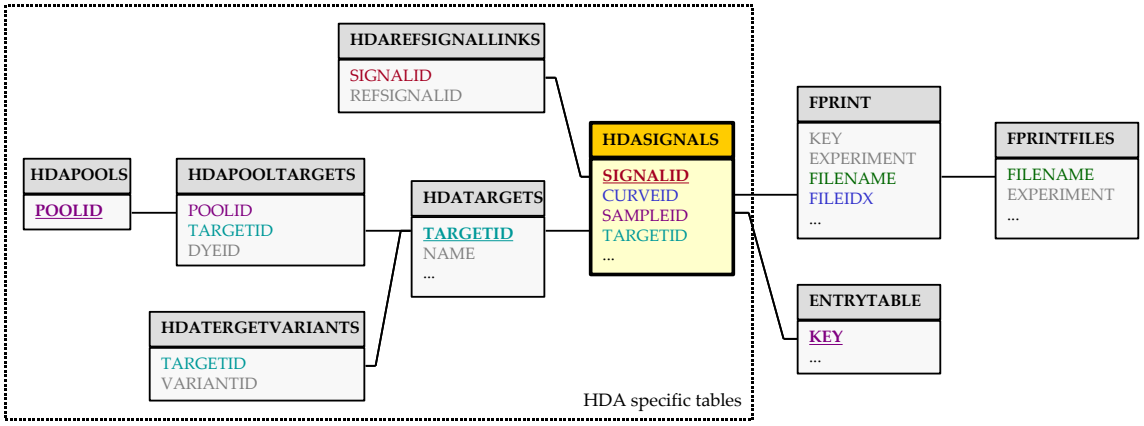

Figure 2.5: Overview of the table structure in a HDA database. Not all columns in the tables are shown.

Chapter 3

Importing HDA profiles

3.1 The BioNumerics main window

The *Main* window consists of a menu, a toolbar for quick access to the most important functions, a status bar, and the following panels (most initially empty) (see Figure 3.1):

- The *Database entries* panel, listing all the available entries (*samples*) in the database, with their information fields and their unique keys.
- The *Experiment types* panel, showing the different experiment types, and the experiments that are defined under each type. In a HDA-configured database, one fingerprint type (**HDA**), and six character types are predefined.
- The *Experiment presence* panel, which for each database entry shows whether an experiment is available (colored dot if available).
- The *Fingerprint files* panel, showing the available data files for an experiment.

The other panels are not of direct use for HDA applications.

If the *HDA plugin* is installed, an extra menu item appears in the main menu: ***HDA***. Most of the functionality needed in relation to HDA is found under this menu item.

3.2 Configuring the pooling strategy

Since each peak signal represents a specific target and belongs to a certain pool, some basic configuration in terms of defining pools and targets needs to be done. It is recommended to have the pools defined prior to importing trace files in the database. The targets, however, can be defined after importing the first batch of trace files. As will be shown later, it is advantageous to define the targets with the densitometric curves shown on the screen, which will facilitate the estimation of a few critical parameters.

2.1 Select ***HDA*** > ***Edit pools***.

This action calls the *Pools* dialog box (see Figure 3.2).

The *Pools* dialog box shows the list of defined pools. Pools can be added (<***Add new***>) and deleted (<***Delete***>). Using the <***Edit***> button targets can be defined that belong to the selected pool.

2.2 Press <***Add new***> and enter a name for the pool, e.g. **Pool1**.

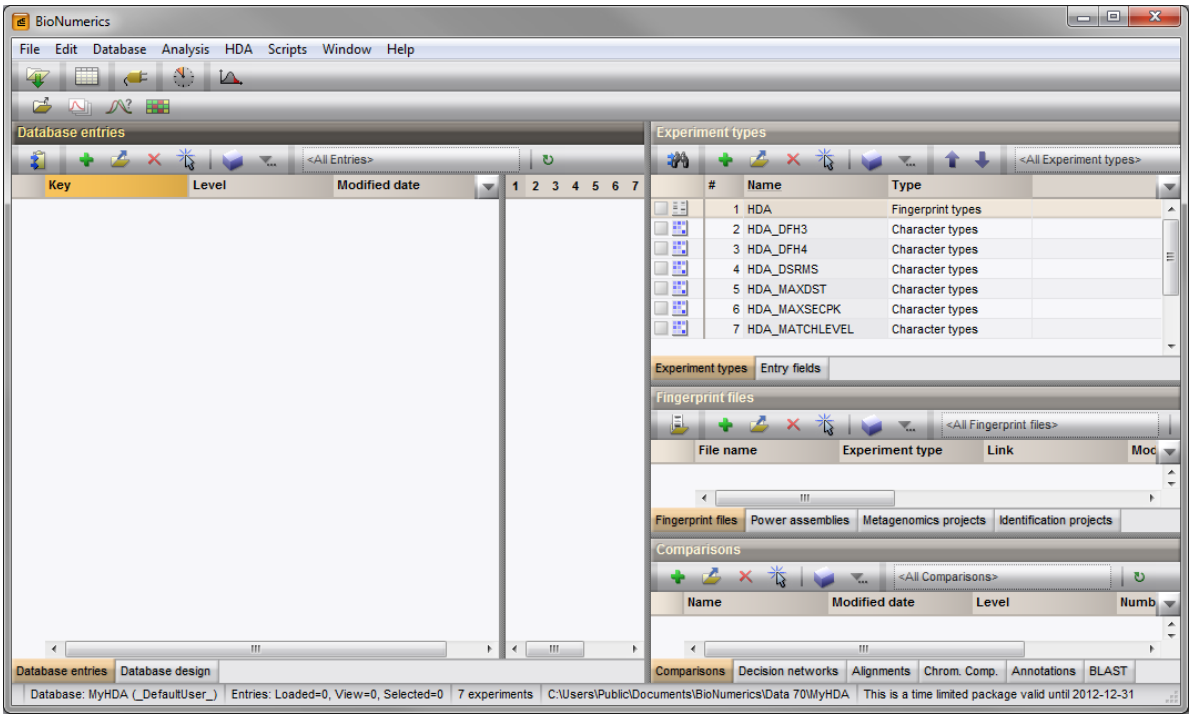

Figure 3.1: The *Main* window after installation of the *HDA plugin*.

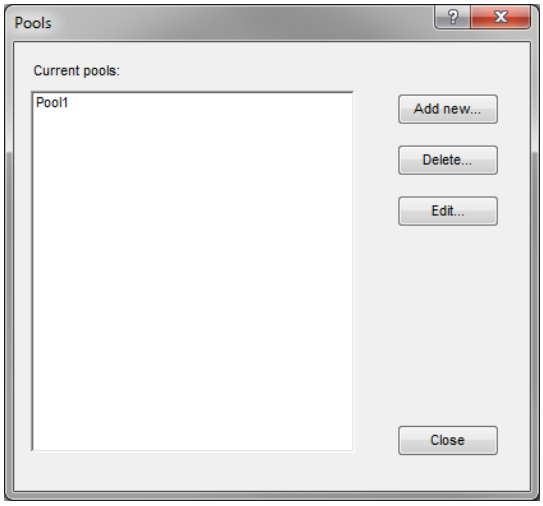

Figure 3.2: The *Pools* dialog box to define and edit pools in a HDA database.

The name should only contain alphanumerical characters with a maximum length of 60 characters. We will define targets after import of the first batch.

2.3 Since there is only one pool available in the present dataset, press the <***Close***> button.

3.3 Importing sequencer trace files

Trace files generated on Applied Biosystems automated sequencers using the GeneScan or GeneMapper software (.fsa files) can be imported directly in BioNumerics. As explained in 2.3, trace files are imported in batch, usually - but not necessarily - corresponding to a run on the sequencer.

As mentioned earlier, the tutorials are based on three sets of 96 hetero-duplex PCR profiles run on a 3100 Genetic Analyzer (Applied Biosystems, Inc., Foster City, CA), obtained by courtesy of Dr. Stephen Edwards, the Institute of Cancer Research, Surrey, UK. The files can be downloaded from the Applied Maths website (`http://www.applied-maths.com/download/sample-data`, click on "CSCE sample data files"). The files can be downloaded as a self-extracting executable, which prompts you to specify a directory to extract the data files. After extraction, the folder contains three sub-folders:

- `FSA_Samples_385_480`
- `FSA_Samples_577_672`
- `FSA_Samples_673_768`

3.1 Select ***HDA*** > ***Import AB trace files***.

3.2 In the file selection box that appears, select all .fsa files in the folder `FSA_Samples_385_480`, and press <***Open***>.

The program now reads and parses the 96 trace files, which may take a while. When finished, the *Import AB densitometric curves* dialog box prompts to enter some information for the fingerprint batch file that will be saved (Figure 3.3).

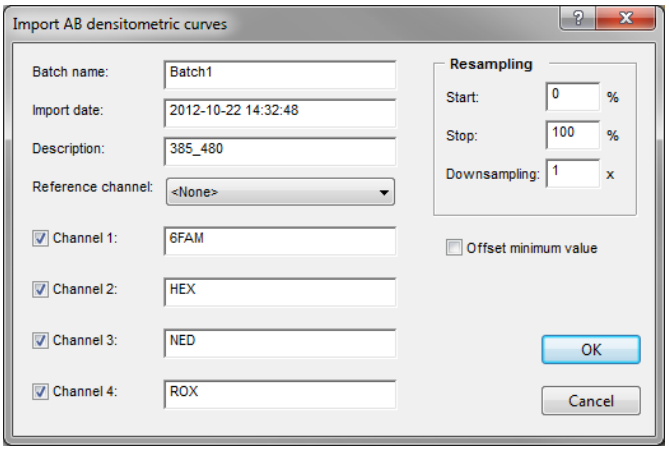

Figure 3.3: The *Import AB densitometric curves* dialog box.

Under ***Batch name***, enter a name for the fingerprint files that will be generated, each containing one dye with 96 traces. The name should only contain alphanumeric characters and hyphens (no spaces, underscores, commas, question marks, etc.). The ***Import date*** is automatically filled in, but can be changed by the user. The ***Description*** field allows a free descriptive text string to be added to the batch of files. Under ***Reference channel***, you can specify one of the dyes present in the trace files as being the reference channel, i.e. containing a set of molecular size markers to normalize the other channels.

The next set of input fields contain the names of the color dyes used. The program reads the dye names from the file and fills them in automatically. The user can overwrite the names if necessary, for example to make different dye nomenclatures compatible. All channels are by default imported in the database (all channel check boxes are checked). Uncheck a check box if you do not wish to import information from a channel into the database.

In the *Resampling panel*, the ***Start*** and ***Stop*** positions are default set to 0% and 100% respectively. Based on these settings, the complete curves are imported. You can change this range if you do not wish to import the complete curves. If the curves have a high resolution, a ***Down sampling*** can be specified to reduce the number of points on the curves. The ***Down sampling*** number is default set to 1 (= no reduction of points).

If the ***Down sampling*** is set to e.g. ”4”, the number of points are reduced 4 times. When the option ***Offset minimum value*** is unchecked, the negative values are clipped to zero. When the option ***Offset minimum value*** is checked, the lowest value is used as offset.

3.3 Enter ”Batch1” as ***Batch name*** and enter the ***Description*** ”385_480”, as specified in the folder name. Since no reference channel is used in the present sample data set, leave this option to <***None***>.

3.4 Make sure all channel check boxes are checked, do not change the settings in the panel on the right and press <***OK***>.

The four channels available in this data set are now being saved as separate files with the name **Batch1** and a suffix _**6FAM**, _**HEX**, _**NED**,and _**ROX**. The saving process may take a while.

When the saving process is finished, the program prompts with the *Determine sample and pool* dialog box to define the name parsing strategy for the densitometric traces (Figure 3.4).

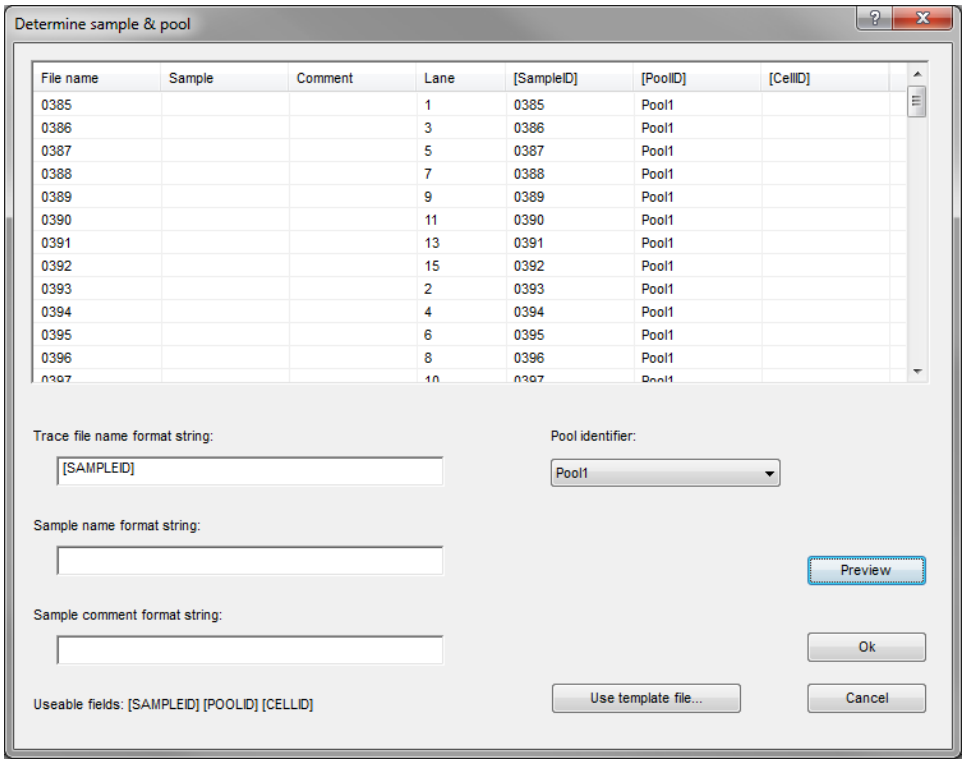

Figure 3.4: The *Determine sample and pool* dialog box to configure the name parsing strategy.

The names of the traces as available in the .fsa files should be chosen so that they contain as much information as possible. This can be done in several ways:

- The file name of the trace file
- The Sample field provided in the trace file
- The Comment field provided in the trace file

The software allows the file name and/or information fields to be parsed based upon components in the name that are separated by a tag. The following components are recognized:

- **[SAMPLEID]**: the name (unique identifier) of the sample

- **[POOLID]**: the pool of fragments run in this trace
- **[CELLID]**: the cell coordinates of a micro plate or other panel that is used to conduct the PCR reactions.

Of these components, only **[SAMPLEID]** is required, the others are optional. A wildcard character (*) can be entered to skip a certain part of the string. Table 3.1 shows some examples of sample names and their corresponding parsing strings.

Sample name	Parsing string
Sample51_Pool3_B11	[SAMPLEID]_[POOLID]_[CELLID]
Sample51|Pool3	[SAMPLEID]|[POOLID]
Sample51	[SAMPLEID]
16-06-2005_Sample51	*_[SAMPLEID]
Sample51(16-06-2005)	[SAMPLEID](*
Sample51(Pool3)_16-06-2005	[SAMPLEID]([POOLID])*

Table 3.1: Examples of sample names and their suggested parsing strings.

Of course, the strings [SAMPLEID], [POOLID] and [CELLID] can be parsed independently from different fields.

To enable the use of different samples within the same trace, import the trace files using a template file in XML format, by pressing the button <***Use template file***>. An example of the template file format is shown below. Please contact Applied Maths for information on this feature.

```
- <CSCETemplate>
  - <Files>
    - <File>
        <Name>A01_0035726085</Name>
        <CellID>A01</CellID>
      - <Signals>
        - <Signal>
            <TargetID>Target001</TargetID>
            <SampleID>Sample001</SampleID>
            <Dye>6-FAM</Dye>
          </Signal>
        - <Signal>
            <TargetID>Target002</TargetID>
            <SampleID>Sample001</SampleID>
            <Dye>VIC</Dye>
          </Signal>
        </Signals>
      </File>
    + <File>
    </Files>
</CSCETemplate>
```

Figure 3.5: Example template file for automated import of trace files without the restriction of having the same sample in one trace file.

3.5 In the present data set, only the sample ID is provided in the trace files. Enter **[SAMPLEID]** under ***Trace file name format string***.

3.6 Since the pool information is not provided in the trace name, specify the pool used by selecting "Pool1" under ***Pool identifier*** (or any other name you entered).

In this set-up, one trace always has to contain fragments from the same sample. One cannot use different samples in different dyes or size ranges within the same trace. To enable the use of different samples within the same trace, you should import the trace files using a template file in XML format, by pressing the button <***Use template file***>.

3.7 Press <***Preview***> to update the preview (optionally) and press <***OK***> (see Figure 3.4).

Pool names may contain up to 60 characters.

At this point, the trace files have been imported in the database. Since this is a first-time import, the targets still need to be defined before the signal processing work can begin.

3.4 Editing dyes

Before proceeding with the analysis of the imported trace files, it is useful to perform some initial database setup. This includes specifying colors and names for the dyes used in the runs.

4.1 Select ***HDA*** > ***Edit dyes*** to call the *Dyes* dialog box (Figure 3.6).

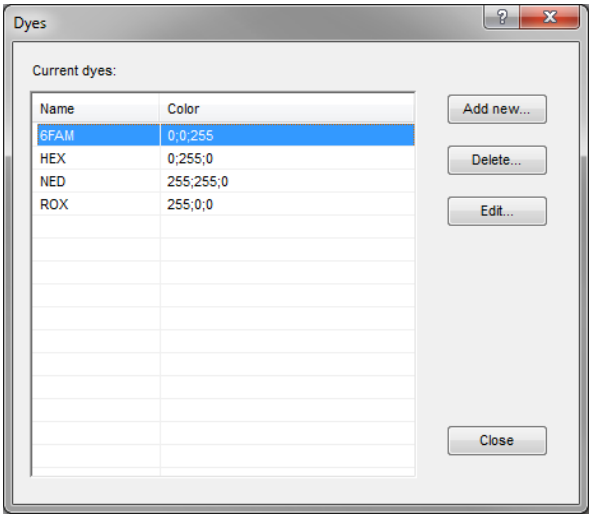

Figure 3.6: The *Dyes* dialog box.

Dyes dialog box lists the dyes that were automatically found by the software. A new dye can be added (<***Add new***>), a selected dye can be deleted (<***Delete***>) and the color components of a selected dye can be defined (<***Edit***>).

4.2 Double-click on **6FAM** or select the dye and press the <***Edit***> button to call the *Dye* dialog box (see Figure 3.7).

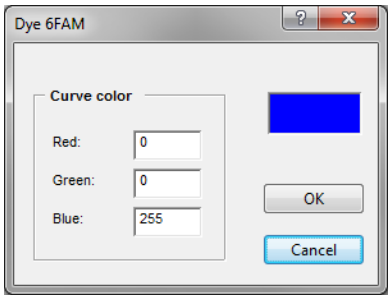

Figure 3.7: The *Dye* dialog box.

In the *Dye* dialog box the Red, Green, and Blue components can be specified. They should range between 0 and 255.

4.3 Enter ”0” for Red and Green and ”255” for Blue, to obtain a blue color for 6FAM.

4.4 Likewise, enter [0,255,0] for HEX (green), [255,255,0] for NED (yellow) and [255,0,0] for ROX (red).

4.5 Press <***Close***> to close the *Dye* dialog box.

3.5 Editing batches of trace files

By selecting ***HDA*** > ***Edit batches*** in the *Main* window, a list of batches present in the database is shown (see Figure 3.8), from which a batch can be selected to perform a number of editing and analysis functions. The descriptive information entered for the batches and their import date is shown as well.

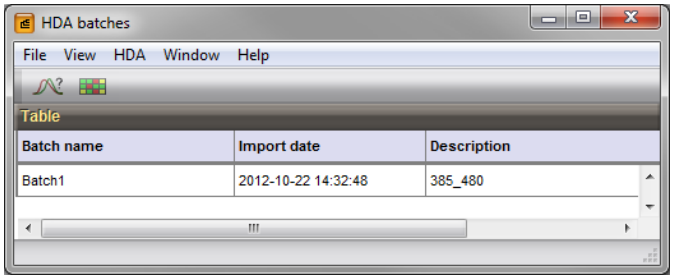

Figure 3.8: The *HDA batches list window*.

From this window, batches can be sorted according to a selected column using ***View*** > ***Sort by column***.

A selected batch can be deleted using ***HDA*** > ***Delete batch***.

5.1 Select **Batch1** and choose ***HDA*** > ***Analyze*** from the menu, or press the button to call the *HDA analysis* dialog box (see Figure 3.9).

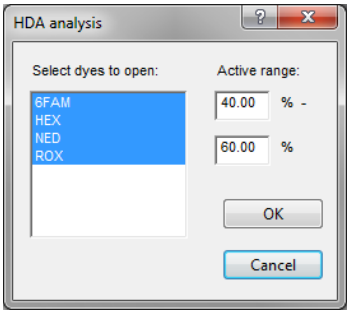

Figure 3.9: Select dyes to open.

The *HDA analysis* dialog box offers the following options:

- Under ***Select dyes to open***, you can either select all dyes, an individual dye, or any combination of dyes to analyze. By default, all dyes are selected.
- With ***Active range***, it is possible to define a range on the curves that will be centered when the *Analysis window* is opened. In addition, the vertical scaling of the densitometric curves will be calculated on the highest value within the specified range. If extremely high peaks occur outside the analysis range, these will not be taken into account so that real signal peaks are vertically scaled to cover the entire range.

The scaling will only be active if you allow the program to scale targets by dye or by lane (see menu option ***Display*** > ***Set active dyes*** in the *Analysis window*).

5.2 Select all four dyes and specify an ***Active range*** between 40% and 60%. Press <***OK***>.

The *HDA curves window* opens in multi-layered view, i.e. with all dyes plotted as overlapping curves (see Figure 3.10).

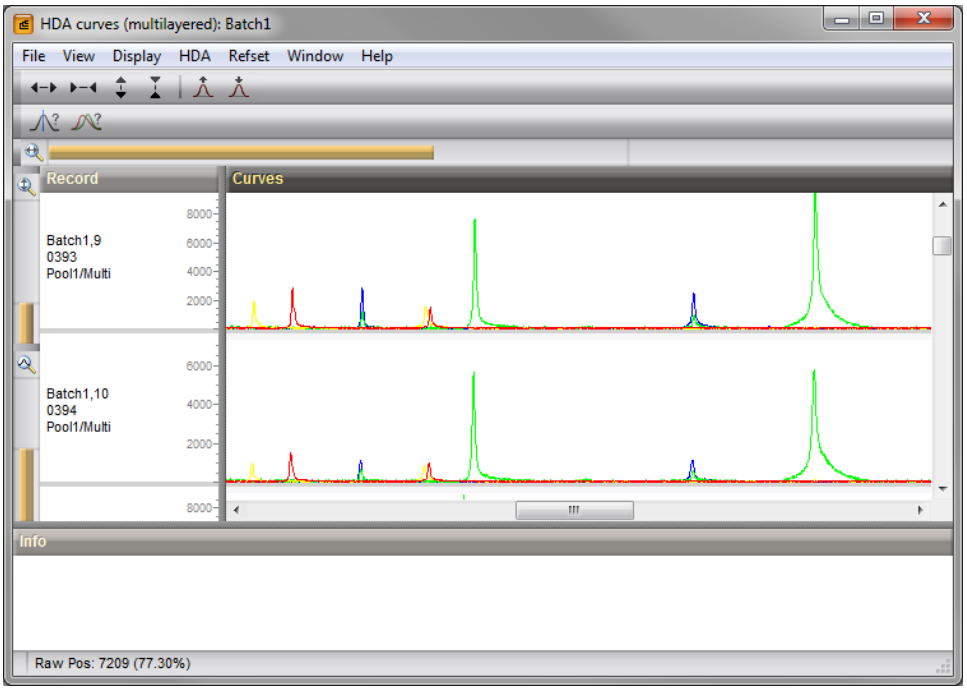

Figure 3.10: The *HDA curves window*: four dyes are displayed in the present example.

The caption of the window indicates the batch. Left from each curve, there are three lines of information:

1. The batch (file) name and the trace number, separated by a comma,
2. The sample name parsed from the trace files,
3. The pool name and dye, separated by a slash.

The dyes are displayed as curves in the colors that were defined earlier (see 3.4). For the inexperienced user, it can perhaps be useful to start working on a single dye rather than displaying all dyes as overlapping curves.

5.3 With ***Display*** > ***Set active dyes***, it is possible to select a subset of dyes to display in the window (see Figure 3.11).

The difference between this option and the initial choice is that all the dyes remain loaded and can be shown again at any time.

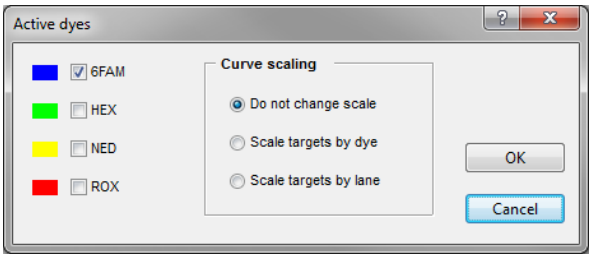

Figure 3.11: The *Active dyes* dialog box.

Any subset of dyes can be selected to be displayed. Using the ***Curve scaling*** option, you can choose how the relative intensities of the traces are scaled. With ***Do not change scale***, the original intensity is maintained. This can be useful to visually inspect the strength of the signals. ***Scale targets by dye*** can be applied to rescale each curve separately according to the highest value found within that dye. When ***Scale targets by lane*** is chosen, the curves are scaled according to the highest value found within one lane, across the dyes. More experienced users, however, can perform the full preprocessing and signal processing with all dyes activated.

If an active range is set (see Figure 3.9), the rescaling will be calculated within the active range.

5.4 For example, uncheck ”HEX”, ”NED” and ”ROX”, to display only ”6FAM”.

5.5 Select ***Scale targets by dye*** and press <***OK***>.

The window now shows ”6FAM” as only dye and the intensities of the curves are rescaled.

The curves can be sorted in different ways using ***Display*** > ***Sort lanes***. This action calls the *Sort lanes* dialog box (see Figure 3.12).

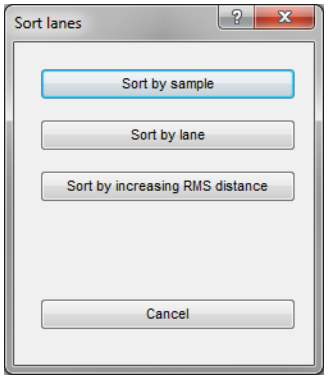

Figure 3.12: Lane sorting options in the *Analysis window*.

The dialog offers following sorting options :

- <***Sort by sample***>: To sort the lanes by sample name.
- <***Sort by lane***>: To sort the lanes by lane number.
- <***Sort by increasing RMS distance***>: This will sort the lanes by increasing correlation to the selected lane (see also 3.16). This function only works after signals have been detected and a matching has been performed.

The *Analysis window* has a number of zoom functions:

- To zoom in on the curve panels horizontally, use ***View*** > ***Zoom in (horizontally)***, in the button bar or drag the zoom slider to the right.
- To zoom out on the curve panels horizontally, use ***View*** > ***Zoom out (horizontally)***, in the button bar or drag the zoom slider to the left.
- To zoom in on the curve panels vertically, use ***View*** > ***Zoom in (vertically)***, click the button in the button bar or use the zoom slider.
- To zoom out on the curve panels vertically use ***View*** > ***Zoom out (vertically)***, click the button in the button bar or use the zoom slider.

- To enlarge the curves use , use ***View*** > ***Enlarge curves*** or the zoom slider.
- To reduce the curves, use , use ***View*** > ***Reduce curves*** or the zoom slider.

The horizontal and vertical zoom settings are saved when the *Analysis window* is closed and will be applied the next time it is opened.

A cursor can be placed on any of the curves. In the status bar of the window, the position of the cursor is given in absolute pixel values and as a percentage of the run length.

In the next paragraphs, the *targets* will be defined and the *signals* on the curves will be interpreted. To keep the work flow in this manual simple and clear, we will work on just one dye. This can be achieved easily with the option ***Display*** > ***Set active dyes*** discussed earlier.

5.6 Make sure "6FAM" is the only dye visualized (see Instruction 5.3). Functions such as signal searching and signal matching (see further) only work on the active (visualized) dyes.

3.6 Defining HDA targets

To define new HDA targets, it is most convenient to have the *HDA curves window* with one dye at a time activated.

6.1 With the *HDA curves window* open and displaying one dye, locate the *Main* window and select ***HDA*** > ***Edit targets*** from its menu to call the *Targets* dialog box (Figure 3.13).

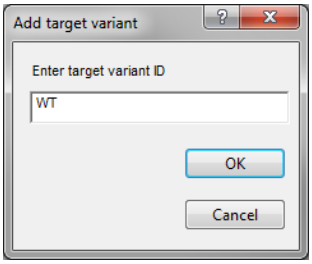

Figure 3.13: The *Targets* dialog box with two targets defined.

The *Targets* dialog box lists the existing targets and allows the user to <***Add new***> targets and <***Delete***> or <***Edit***> existing targets. Pressing the <***Copy settings***> button calls the *Copy target settings* dialog box where parameter settings can be copied from one target onto another.

6.2 Press <***Add new***> to enter a target.

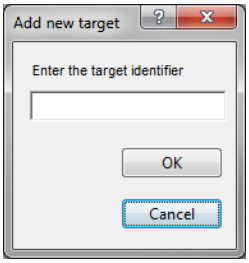

Figure 3.14: The *Add new target* dialog box.

The *Add new target* dialog box prompts for a target name. The name should only contain alphanumerical characters (e.g. "Target1Dye1").

6.3 Enter a name for the target, e.g. "Target1Dye1" and press <***OK***>.

The target is added to the *Targets* dialog box.

6.4 Double-click on the newly created target to call the *Target* dialog box where the target settings can be defined (see Figure 3.15).

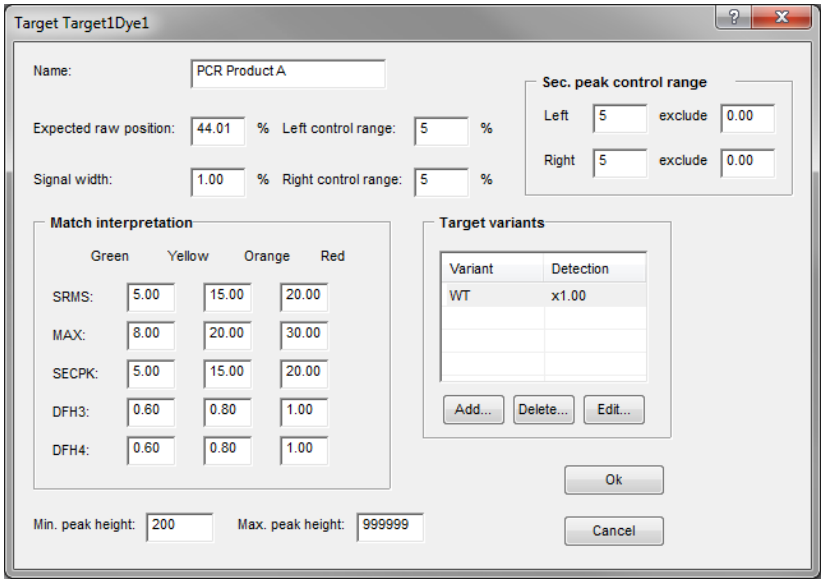

Figure 3.15: The *Target* dialog box: PCR Product A.

In the *Target* dialog box following settings can be entered :

- Under ***Name***, a descriptive name can be entered for the target. This name can include spaces and any other characters.
- The ***Expected raw position*** is the approximate position of a signal for that target, as a percentage of the run length.
- The ***Signal width*** is the approximate width of the peak or peaks that constitute the signal, as a percentage of the run length.
- The ***Left control range*** is an area left from the signal within which the algorithm will look for additional peaks that might constitute signal from mutations.
- The ***Right control range*** is the area right from the signal within which the algorithm will look for additional peaks that might constitute signal from mutations.
- The ***Left secondary peak control range*** is the area left from the signal within which the algorithm will look for secondary peaks. The ***Exclude*** range is the area left from the signal that will be excluded from the search for secondary peaks.
- The ***Right secondary peak control range*** is the area right from the signal within which the algorithm will look for secondary peaks. The ***Exclude*** range is the area right from the signal that will be excluded from the search for secondary peaks.
- The ***Match interpretation parameters*** are threshold values for the interpretation of five different peak matching parameters (SRMS, MAX, SECPK, DFH3, and DFH4). The meaning of these parameters and setting of the values are discussed in 3.16.
- ***Minimum peak height*** is a threshold that allows the program to ignore signal that is too weak.

- Similarly, ***Maximum peak height*** is a threshold that allows the program to ignore signal that is too strong. By default, this parameter is set to ”999999”. If you enter zero or leave the field empty, the program will interpret the value as infinite height, and translate it into ”999999”.
- With ***Target variants***, different, known ”normal” polymorphisms for the same target can be defined (see below).

6.5 Enter for example ”PCR Product A” as descriptive name for the target.

In the present sample set, targets with two different lengths have been pooled in one trace and using the same dye, which means that each individual curve is composed of two signal peaks. It is therefore important to define the expected raw position and the left and right control ranges with some care. The ***Expected raw position*** can easily be derived from the *HDA curves window*, as follows:

6.6 In the *HDA curves window*, click on the left-most of the two signal peaks in one of the traces. Choose a WT trace with just one single, well formed peak without shoulders, and remember which trace you chose. The exact position, as a percentage, is indicated in the status bar of the window (it should fall between 44% and 47% for the left signal peak). If large shifts occur between the peak positions of the traces, make sure to use an average trace in terms of the position of the peak.

If you do not use the first trace, remember the trace number to be able to use the same trace to add the second target.

6.7 In the *Target* dialog box, enter the position under ***Expected raw position*** (e.g. ”44.81”).

6.8 The same procedure can be repeated to determine the ***Signal width***. Make sure to enter a width that is large enough to cover composite signal, e.g. composed of 4 peaks such as found in heterozygous polymorphic samples. A suggested width for this sample set is 1%.

6.9 For the ***Left*** and ***Right control ranges***, enter two ranges that are as small as possible, yet including any possible secondary peaks that are relevant to the interpretation of the curves, but excluding irrelevant signal such as occurring at the front of the curves. A suggested control range for the sample data set is 5% at either side of the peak position.

If the right control range of the left signal overlaps with the left control range of the right signal, the average position between both will be taken by the program as limits.

6.10 Enter a left and right ***Secondary peak control range*** of 5%.

6.11 As ***Target variants***, at least one variant has to be entered to make the further analysis possible. Press <***Add***> to call the *Add target variant* dialog box.

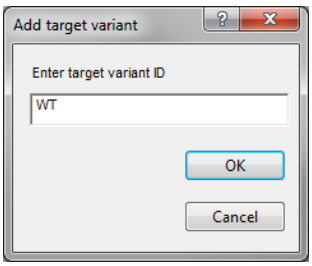

Figure 3.16: The *Add target variant* dialog box.

The *Add target variant* dialog box prompts for the target variant name.

6.12 In this example, enter ”WT” as a target variant name and press <***OK***>. In other words, we assume that there is only a wild type available but no polymorphisms.

When added, the target variant WT appears with a ***Detection factor*** of x1.00.

The detection factor determines the sensitivity of the detection algorithm by multiplying the result found after comparison of the signal with the reference(s) with the specified factor. If the factor is less than 1, for example 0.5, the distances found between the signal and its closest reference will be multiplied by 0.5, resulting in a smaller distance and easier acceptance. Conversely, a detection factor higher than 1, e.g. 3, will cause the distances to be multiplied by 3, resulting in higher distances and less easy acceptance.

The default detection factor of 1 should not normally be changed for the ”wild type” variant of the target. However, in case of known polymorphisms with multiple peaks and/or shoulders, it might be necessary to specify a detection factor of less than 1 for those variants to allow the software to find them correctly. To do so, you should double-click on the target variant or press the <***Edit***> button (not needed in this example).

6.13 Press <***OK***> to save the target.

6.14 Repeat the whole procedure starting from step Instruction 6.2 to add a second target, corresponding to the second peak in the same dye. This peak should fall between 52% and 57%. To define the ***Expected raw position***, use the same trace as the one used for defining the first target (see Instruction 6.6). It is suggested to enter ”Target2Dye1” as identifier, and in the *Target* dialog box specify ”PCR Product B” as name, ”1”% as signal width, and ”2”% as left and right control ranges. Enter a left and right ***Secondary peak control range*** of ”5”%. One target variant, ”WT” also has to be defined.

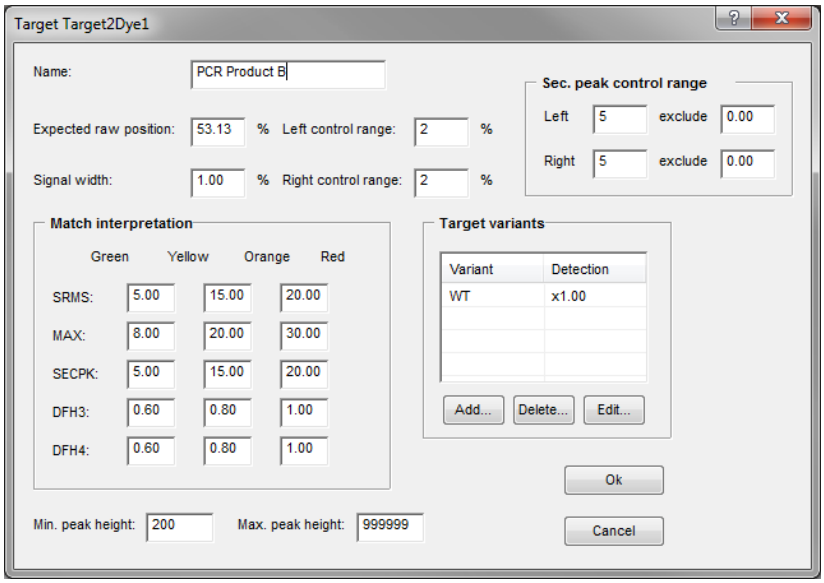

Figure 3.17: The *Target* dialog box: PCR Product B.

6.15 Close the *Targets* dialog box.

3.7 Linking targets to pools

Before the program can actually start processing the first batch, it needs to know what targets are found in the defined pools.

7.1 In the *Main* window, select ***HDA*** > ***Edit pools*** from the main menu to call the *Pools* dialog box.

7.2 Select ”Pool1” and press <***Edit***> to call the *Pool* dialog box.

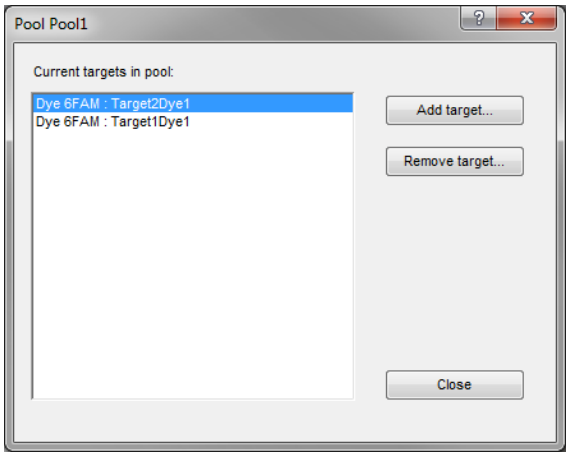

Figure 3.18: The *Pool* dialog box: Two targets are defined for "Pool1".

In the *Pool* dialog box targets can be added (<***Add target***>) or removed (<***Remove target***>) from the selected pool.

7.3 Press <***Add target***> to call the *Add target to pool* dialog box (see Figure 3.19).

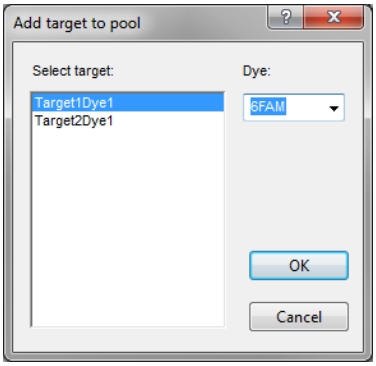

Figure 3.19: The *Add target to pool* dialog box.

In *Add target to pool* dialog box one needs to select the target and corresponding dye that should be found in the selected pool.

7.4 Select "Target1Dye1" and dye "6FAM" and press <***OK***>.

7.5 Press <***Add target***> again, select "Target2Dye1" and dye "6FAM" and press <***OK***>.

7.6 If the two targets are defined within the pool (both should be listed along with the dye number, see Figure 3.18), press <***Close***> to finish editing the pool.

7.7 Press <***Close***> to close the *Pools* dialog box as well.

The database is now sufficiently configured to start processing batches of HDA profiles. It is still possible to modify the parameters for targets, or to add/edit targets and pools at any time.

3.8 Finding signals on HDA profiles

8.1 If the *HDA curves window* for **Batch1** is not opened, select it in the HDA batches list and choose ***HDA*** > ***Analyze***.

8.2 In the *HDA curves window*, select ***HDA*** > ***Find targets***. Confirm the action.

The program now searches for the positions of the target peaks of the *active dyes* based upon the approximate positions that were entered in the *Target properties dialog box*. When finished searching, a dialog box reports the errors encountered. In the sample set used in this example (dye 6FAM), target signal is missing in lanes 95 and 96. The result is shown in Figure 3.20. Each target signal found is highlighted with a gray bar, the width of the highlight zone corresponding to the expected signal width for the target. A brown or blue bar indicates the control range specified for the target. For each signal found, the target identifier to which it has been assigned is indicated.

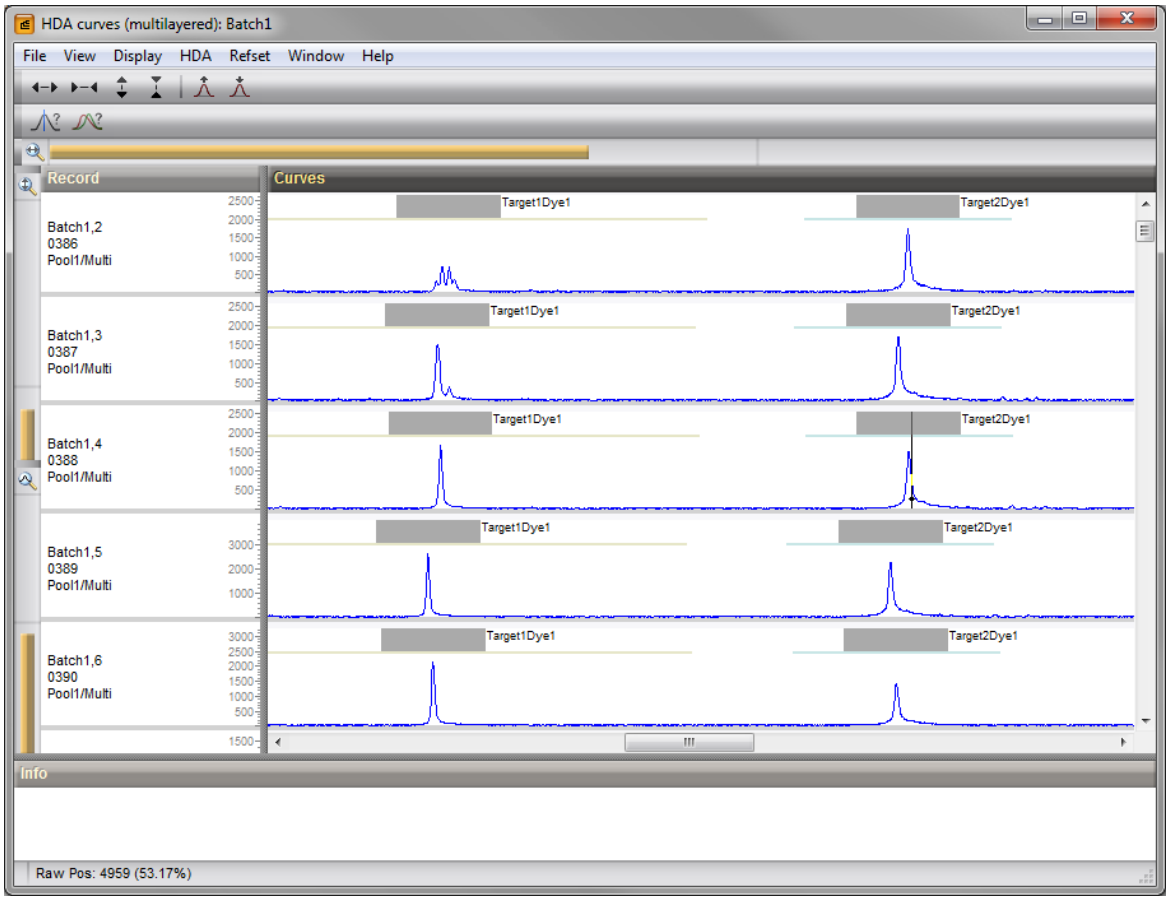

Figure 3.20: The *HDA curves window* after searching for target signals.

The vertical position of the gray bar is different depending on the dye selected, so that the bars do not overlap in case more than one dye is active.

8.3 A signal can be removed by clicking within the gray highlight zone and pressing the **Del**-key (or ***HDA*** > ***Remove signal***).

A new signal can be added by specifying the left and right control ranges on the curve:

8.4 Click on the left limit of the target range and hold down the mouse button.

8.5 Drag the mouse pointer to the right limit of the target range and release it there.

8.6 Two vertical bars now appear on the curve. To add a signal within this range, press the **Insert**-key or choose ***HDA*** > ***Add signal***.

3.9 Matching signals using references

Once signals are found on the HDA profiles, a matching can be done of all signals against a defined set of reference signals. The reference signals are usually assigned by the user and should be based upon profiles that are known to belong to a wild type or a polymorphic type.

The user can either define reference signals on every new batch of HDA profiles that is analyzed, or create *Reference sets*, which are stored sets of reference signals that can be chosen to use as references for any new batch of HDA profiles that is analyzed (see 3.15).

9.1 For this example, choose a normal-looking signal from **Target1**, for example curve 5.

9.2 To mark the signal as a reference, click on the signal (gray) bar and press the **Enter**-key (or ***HDA*** > ***Edit signal***, or double-click) to call the *Edit signal* dialog box (see Figure 3.21).

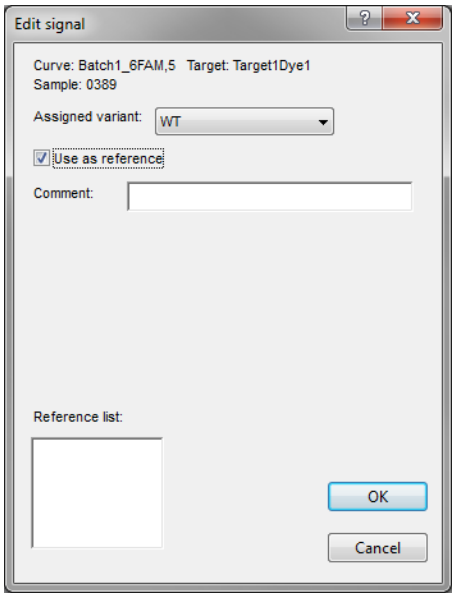

Figure 3.21: The *Edit signal* dialog box.

The *Edit signal* dialog box allows the user to mark the target signal as belonging to one of the specified polymorphic variants (e.g. "WT") with ***Assigned variants***. With ***Use as reference***, the signal can be marked as a reference signal and will be used to match the others against. Furthermore, a ***Comment*** line can be added for the signal. The ***Reference list*** will display the references against which a selected non-reference signal has been matched.

9.3 In the *Edit signal* dialog box, select ***WT*** as ***Assigned variant*** and check ***Use as reference***.

9.4 Press <***OK***> to confirm the editing and to close the dialog box.

A signal that is marked as a reference is indicated with a blue bar. The polymorphic type (WT or variant) is indicated right from the signal.

9.5 Repeat Instruction 9.1 to Instruction 9.4 for a normal looking signal of **Target2**, e.g. lane 5 as well.

It is possible to add any number of references if these are known. However, the processing time of the matching algorithm is in proportion to the number of reference signals defined. It is therefore not recommended to define more references than necessary to reflect the natural variance in the signals and to obtain an appropriate matching.

9.6 To perform the matching, select ***HDA*** > ***Perform match***. This calls the *Match HDA signals* dialog box (see Figure 3.23).

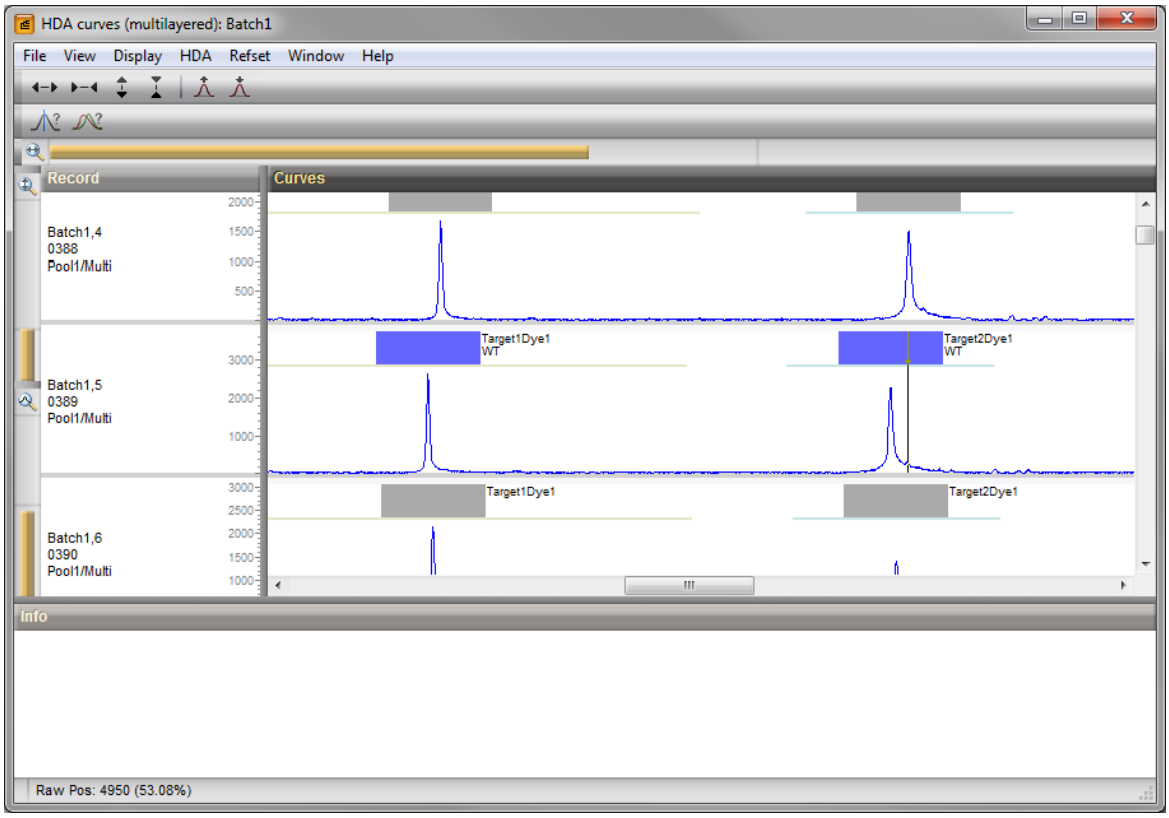

Figure 3.22: Signals marked as reference are marked with a blue bar.

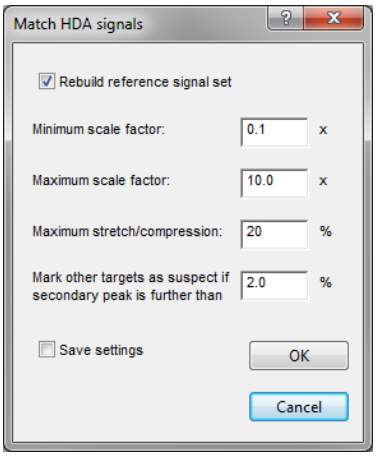

Figure 3.23: The *Match HDA signals* dialog box with HDA matching settings.

The *Match HDA signals* dialog box prompts for a number of settings:

Before calculating the matching parameters, the algorithm performs a physical matching of the peaks with the reference peaks. This involves a stretching or shrinking in the horizontal direction and a scaling in the vertical direction.

- The ***Minimum scale factor*** and ***Maximum scale factor*** can be entered and have default values of 0.1x and 10x, respectively.
- The ***Maximum stretch/compression*** factor can also be set and is 20% by default.
- With ***Mark other targets as suspect if secondary peak is further than***, it is possible to provide an

alerting tool for peaks that have uncertain origin, i.e. that could be resulting from two targets present in the same pool and dye. If only one target is present in a curve, this alerting tool will be ignored by the program. In addition, secondary peaks will only be marked as suspect if they occur *between* two targets on the curve.

- ***Rebuild reference signal set*** ensures that the information on reference signals is updated before the matching starts. For each signal, the software stores the set of reference signals that was used for the previous matching. If the ***Rebuild reference signal set*** check box is not checked, this previous reference signal set will be used (if any) regardless of whether changes to the reference signals were made or not. If no previous reference signal set exists for a signal, a new set will be built based on the most recent information available.

When checking the option ***Save settings*** the settings are saved as the new default settings in the database.

9.7 Enter ”4”% for the ***Mark other targets as suspect*** parameter and press <***OK***> to start the matching.

The matching is executed only on the currently active dye(s).

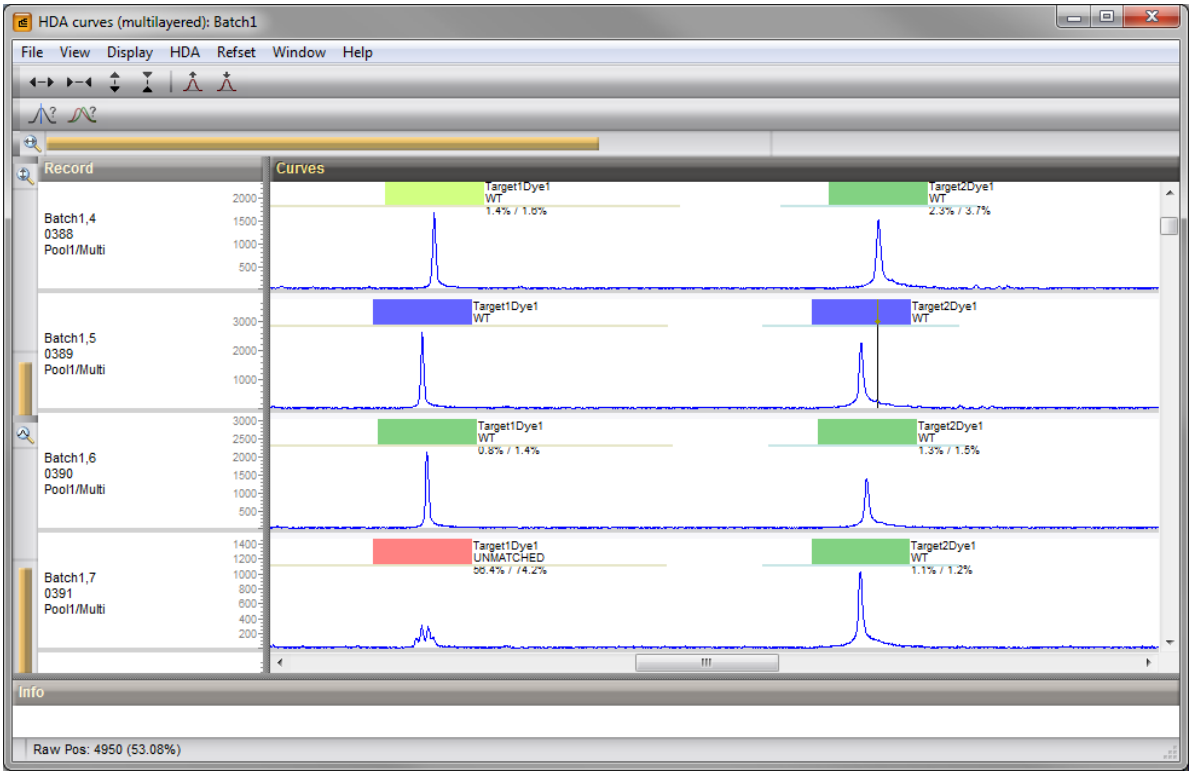

Figure 3.24: The *HDA curves window* after performing a matching.

When finished, each signal has a colored background (see Figure 3.24), of which the meaning is as follows:

- **Blue**: Signal is defined as reference.
- **Green**: Signal shows a very good match with at least one reference. Signal is identified as a wild type or polymorphic variant.
- **Yellow**: Signal has a fairly good match with at least one reference. Signal is identified as a wild type or polymorphic variant.
- **Orange**: Signal has a weak but insufficient match with the reference(s). Signal is labeled as UNMATCHED.

- **Red**: Signal does not match any of the references. Signal is labeled as UNMATCHED.
- **Gray**: The signal failed to pass the criteria applied to the target, more specifically, the minimum peak height (see 3.6). Signal is labeled as FAILED.

If one signal in a curve is marked as failed, the other signal (if any) is marked with a red bar, an indication that this curve is suspect. If there is a secondary peak from a signal in the curve that falls beyond the maximum percentage specified in Instruction 9.7, the other signal (if any) will be marked as suspect too.

3.10 Matching display and comparison tools

After matching, the polymorphic variant to which the signal has been assigned is indicated on the second line right from the signal. The third line shows a percentage deviation from the closest reference using the first two (and most important) parameters used by the matching algorithm: ***SRMS*** and ***MAX***. The definition of the matching parameters and the way to adjust them are explained in 3.16.

If it appears that, after matching, some wild types or non-mutant polymorphic variants remain unmatched, one can add one or more examples of those signals as additional references and rerun the matching algorithm.

10.1 For a selected signal (click within its colored bar), it is possible to display a detailed comparison with the closest reference by pressing the **Ctrl+Enter** on the keyboard (or ***HDA*** > ***Show signal match***).

The result, where two curves are plotted onto each other,is shown in Figure 3.25.

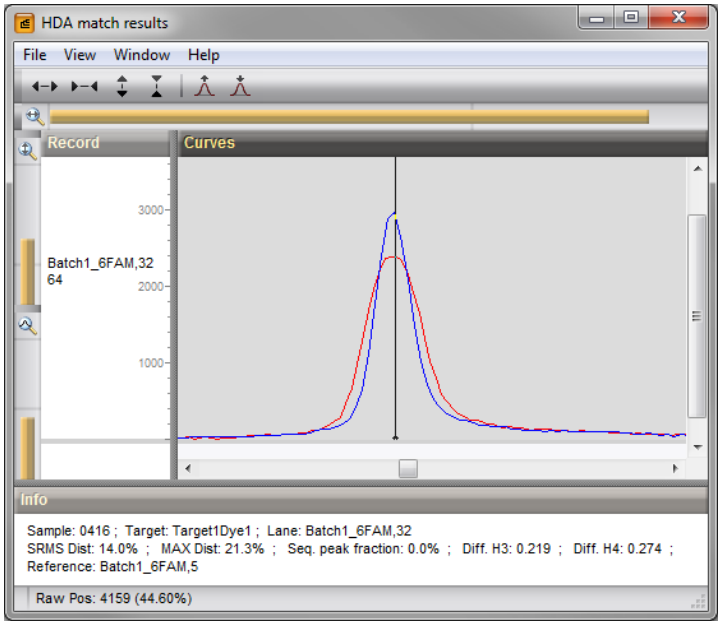

Figure 3.25: Detailed signal to reference comparison.

The reference curve is shown in blue; the signal curve in red. The same zoom functions as described for the *HDA curve window* apply to this window. The window displays the information for the selected sample, target and lane and a full overview of the results from the 5 matching parameters (see 3.16). The closest reference used to compare is also indicated.

You can open several detailed comparison windows simultaneously to compare side by side.

10.2 Close the detailed comparison window to return to the *HDA curves window*.

10.3 With the cursor placed on a signal, it is possible to align all corresponding signals vertically with ***Display*** > ***Align signals***. In case multiple targets are present in one dye (like in the example data set), only the selected signal will be aligned.

10.4 Another useful feature is ***Display*** > ***Sort lanes*** (see also Figure 3.12). Pressing the button <***Sort by increasing RMS distance***> will sort the signals on the basis of overall signal resemblance with the closest reference signal.

3.11 Matching editing tools

A matching result automatically assigned by the matching algorithm can be manually overruled by the user.

11.1 As an example, select a signal on a red or orange background, which is labeled as **UNMATCHED**.

11.2 Press the **Enter**-key to open the *Edit signal box*. The box now looks as in Figure 3.26.

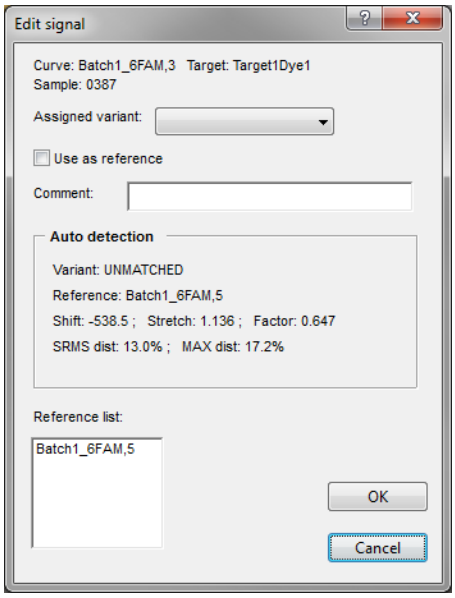

Figure 3.26: The *Edit signal box* after automatic matching.

Under ***Auto detection***, the result from the automatic matching is displayed: the variant (unmatched), the closest reference used, and the shift, stretch and scaling factors applied to match the signal. The two most important matching parameters **SRMS** and **MAX** (see 3.16) are also indicated.

Reference list displays a list of the reference signals to which the actual signal has been matched.

11.3 Click the ***Assigned variant*** drop-down list and select ”WT”.

11.4 Press <***OK***> to confirm the change. The signal is now marked as **WT** and has a pale blue background, which indicates that it has been assigned manually.

If a signal is defined manually, it will not be overwritten during subsequent automatic matching cycles. To undo a manual assignment, open the *Edit signal box* and select the empty field as ***Assigned variant***.

If ***HDA*** > ***Find targets*** is selected after a matching and/or manual assignment, all existing information on automatic assignments, manual assignments, and reference signals will be lost!

3.12 The multi–layered analysis view

In the multi-layered analysis view, the trace curves from different dyes within the same lane are displayed as overlapping curves (see also 3.5). This mode can be useful not only to achieve more information in a condensed view, but also to compare the positions of peaks to positions on another dye, for example a dedicated set of reference markers. The multi-layered view will probably be the preferred mode by more experienced users.

12.1 Open the Batch editor with ***HDA*** > ***Edit batches*** (see 3.5).

12.2 Select "Batch1" and choose ***HDA*** > ***Analyze*** from the menu in the *Batch editor window*.

12.3 In the *HDA analysis dialog box* (see Figure 3.9), select all four dyes (default).

12.4 As ***Active range***, set 40% - 60%, and press <***OK***>.

All dyes are now displayed and are overlapping within the same lane (see Figure 3.27 for an example).

Figure 3.27: The *Analysis window*, multi-layered mode.

3.13 Cloning of target settings

Defining a new target requires a number of parameters to be filled in correctly. It may therefore be useful to be able to copy all settings from one target onto another. This can be achieved in the *Target selection dialog box* (see Figure 3.13), which is called from the *Main* window with ***HDA*** > ***Edit targets***.

13.1 If you do not want to copy the settings onto an already existing target, first create a new target (Figure 3.13): see Instruction 6.2 to Instruction 6.3.

13.2 Press the <***Copy settings***> button to call the *Copy target settings* dialog box (see Figure 3.28).

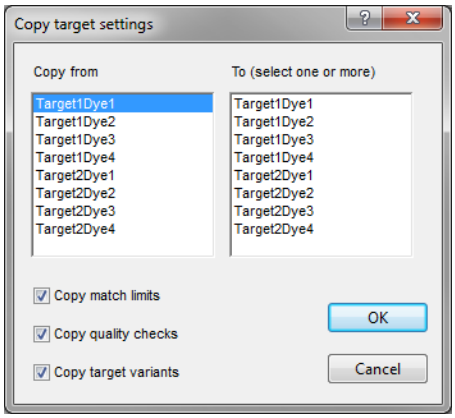

Figure 3.28: The *Copy target settings* dialog box.

The dialog displays two list boxes containing the existing targets. Select the target to ***Copy from*** in the left list box, and the target to ***Copy to*** in the right list box. With the **Ctrl**-key more than one target can be selected.

Be careful with copying settings to existing targets, as the original settings for the target will be lost!

Optionally, the tool allows you to copy the matching parameter settings (***Copy match limits***), the minimum and maximum peak height settings(***Copy quality checks***) and the defined target variants for the source target (***Copy target variants***). By default, these three options are checked.

The name, the expected raw position and the left/right control ranges are not copied as these are believed to be unique per target.

3.14 Automatic detection of references

The *HDA curves window* provides a tool to detect suitable reference signals automatically. The tool is based on the assumption that wild-type signals and their polymorphic forms constitute a significant fraction of the total number of signals in the batch that exhibit very similar profiles.

14.1 To have the software find reference signals automatically, select ***HDA*** > ***Auto detect references*** in the *HDA curves window*.

When the search is finished, a number of signals are marked as references.

3.15 Working with stored reference sets

It is possible to store appropriate sets of reference signals in the database as *Reference sets*, so that they can be re-used for every new batch of HDA profiles that is analyzed. A condition for the successful use of such reference sets is that the reproducibility of the technique is so high that inter-run variability between HDA profiles is not noticeably higher than within-run variability.

15.1 In the *Main* window, select ***HDA*** > ***Edit reference sets*** to call the *Reference sets* dialog box (see Figure 3.29).

In the *Reference sets* dialog box you can ***Add*** a reference new set, ***Edit*** or ***Delete*** an existing reference set.

15.2 Press <***Add***> to call the *Add reference set* dialog box (see Figure 3.30).

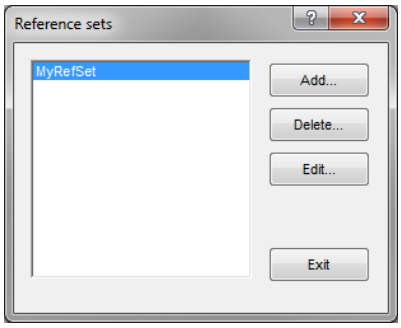

Figure 3.29: The *Reference sets* dialog box to create and edit reference sets.

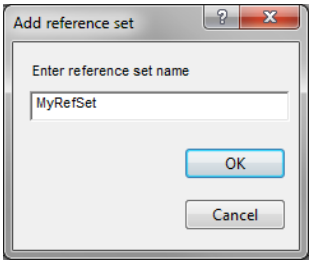

Figure 3.30: The *Add reference set* dialog box.

The *Add reference set* dialog box prompts for a reference set name.

15.3 Enter a name, e.g. ”MyRefSet”, and press <***OK***>.

15.4 Close the *Reference sets* dialog box with <***Exit***>.

15.5 Open the *HDA curves window* of **Batch1**.

15.6 Make sure that reference signals are defined for this batch (blue bars), and if so, choose ***Refset*** > ***Add all references to set*** to call the *Add all references to set* dialog box (see Figure 3.31).

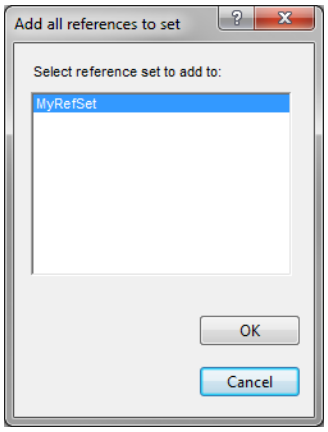

Figure 3.31: The *Add all references to set* dialog box to populate a reference set.

In the *Add all references to set* dialog box the reference sets appear in a list box. Pressing <***OK***> adds the reference signals from the batch to the reference set.

15.7 Press <***OK***> to add the reference signals.

The reference set can be further populated with references from other batches as well, or can be edited as follows:

15.8 Close the *HDA curves window*.

15.9 In the *Main* window, choose ***HDA*** > ***Edit reference sets***.

15.10 Select the created reference set (**MyRefSet**) and press <***Edit***> to call the *Reference set* dialog box (see Figure 3.32).

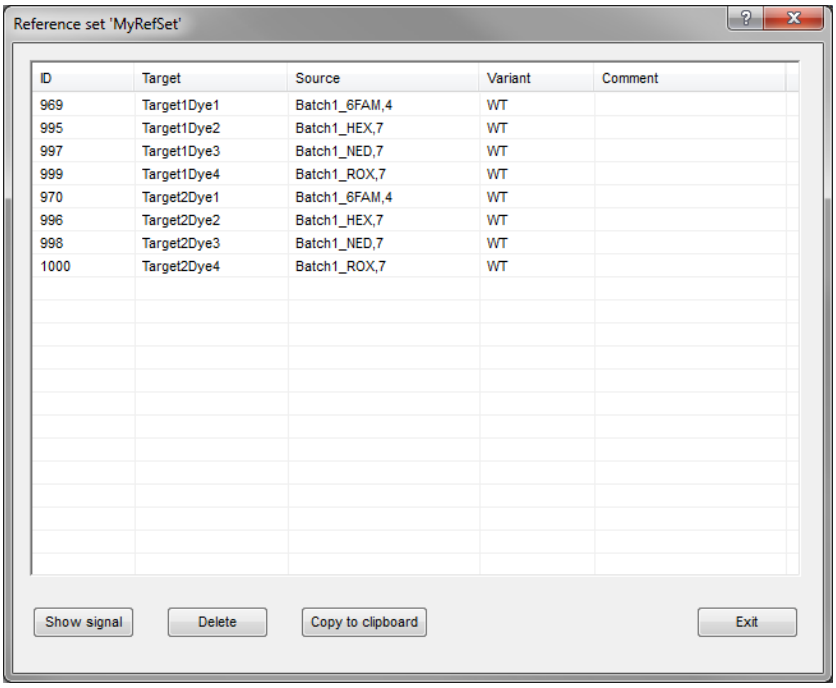

Figure 3.32: The *Reference set* dialog box.

Following functions are available:

- <***Show signal***> displays the detailed signal curve window for the selected reference signal from the list.
- <***Delete***> removes the selected signal(s) from the list (note that the list supports multi-line selection).
- <***Copy to clipboard***> copies the list to the Windows clipboard as a tab-delimited text file.

3.16 Calibrating the matching parameters

For its HDA signal matching algorithm, BioNumerics uses a number of parameters for which the threshold values can be modified by the user. These values can be changed per target in the *Target* dialog box (see 3.6). Below are the parameters and their meaning.

- **SRMS**: The Single-sided Root Mean Square value, which is a root-mean-square value calculated on the highest 50% of square differences between the curves compared. The single-sided modification ensures that small background noise is not included as difference between the curves.
- **MAX**: This parameter simply determines the highest difference between any corresponding values on the two curves, after some smoothing.

- **SECPK**: Calculates the proportion of the secondary peak to the primary of the curve.
- **DFH3**: For each peak on a curve, the deviation from a Gaussian is calculated. H3 is a measure of the deviation from symmetry. DFH3 is the difference in H3 parameters between the signal and the reference. If the signal has the same shape as the reference, this difference is supposed to be zero.
- **DFH4**: A similar parameter as DFH3. H4 determines the deviation from a theoretical Gaussian, and DFH4 is the difference between the H4 parameters from the signal and the reference peaks.

The first two parameters SRMS and ***MAX*** are the only parameters that take into account **any** difference between the curves, and that could thus be used as valid standalone estimators of matching. The other parameters, SECPK, DFH3, and DFH4 each apply to a specific component of the curve (the peak shape, the secondary peak).

To facilitate the estimation of the right parameter values for a given target, the software can show a calibration window, showing a histogram of matches and mismatches found per parameter.

16.1 From the *HDA batch list window* (***HDA*** > ***Edit batches***; see 3.5), select a batch and choose ***HDA*** > ***Calibrate*** to call the *HDA calibration* dialog box (see Figure 3.33).

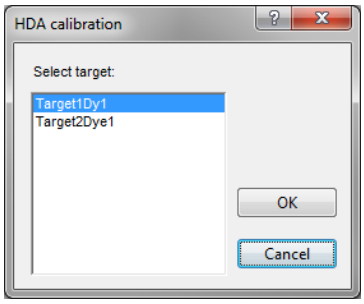

Figure 3.33: The *HDA calibration* dialog box.

The dialog prompts you to choose a target from the list of available targets for the batch.

16.2 Select for example "Target1Dye1" and press <***OK***>. The result is shown in Figure 3.34.

Each histogram shows the distribution of matches and mismatches as obtained from a parameter used by the software. The green bars indicate signals that were found matching to a reference under the current settings or that were manually set to be matching by the user. The red bars indicate signals that could not be matched to a reference under the current settings or that were manually set as unmatched by the user. If part of a bar is brown, it means that either a red or a green smaller bar is shown on top of a bar of the other color.

Based upon the distribution found for each signal, the software proposes an optimal set of threshold values between the green-yellow-orange-red coloring in the matching window. The values are also indicated on top of each histogram.

16.3 These or any other values can be compared to the current settings for the target in the *Target dlg* (see 3.6).

Any of the proposed threshold values can be changed manually as follows:

16.4 Select with the mouse a location in the histogram of the parameter to change: a vertical pointer line appears.

16.5 Select ***HDA*** > ***Edit limit*** to call the *Edit calibration limit* dialog box (see Figure 3.35).

In the *Edit calibration limit* dialog box you can specify the limit and the value for the selected parameter. The value that is entered automatically corresponds to the location of the pointer on the histogram.

16.6 Press <***OK***> to save the new threshold value or <***Cancel***> to exit without saving.

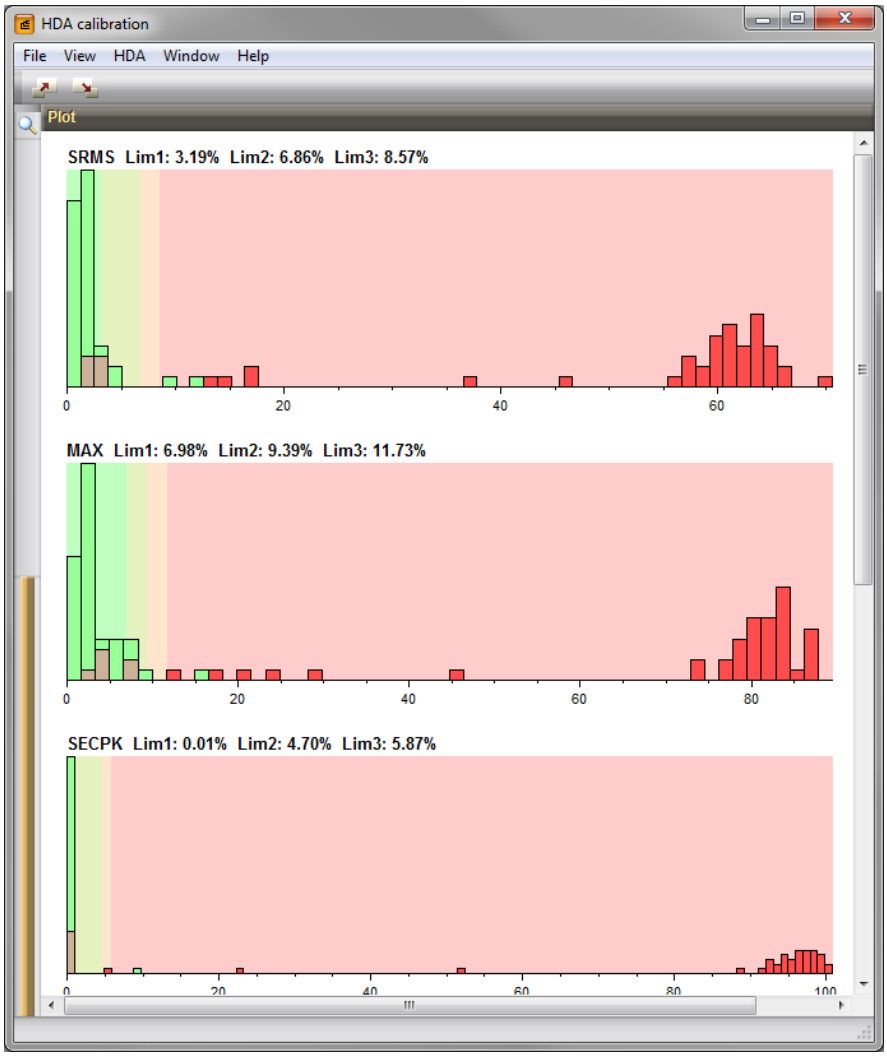

Figure 3.34: The *Parameter calibration histogram window*.

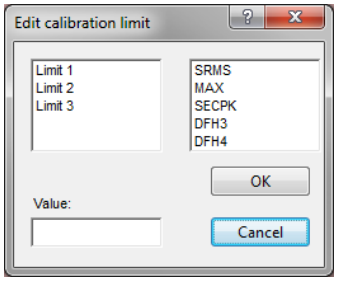

Figure 3.35: The *Edit calibration limit* dialog box to enter parameter thresholds from the calibration window.

16.7 You can have the software copy the proposed settings automatically to the target properties by selecting ***HDA*** > ***Save new limits***.

To view the results of a matching after changing the parameter values, perform a new matching in the *HDA curves window*.

16.8 The calibration report can be printed using ***File*** > ***Print***, or an enhanced metafile can be copied to the clipboard with ***File*** > ***Copy to clipboard***.

The software uses the most strict criterion possible to mark a signal as mismatched: if any of the 5 parameters exceeds the mismatch threshold, the signal is marked as a mismatch.

It is possible, for a mismatched signal, to inspect which parameter is responsible by clicking **Ctrl+Enter** in the *HDA Curves window* to pop up the detailed comparison. The values found for each parameter are listed in this window, and can thus be compared to the limits specified for the target.

It is possible to exclude a parameter for use by the matching algorithm by specifying very high limits, such as for example 200 for each limit.

The calibration tool takes into account any corrections made to the automatic matching by the user. For example, if the user has changed a number of signals originally found as mismatched into matches, it will propose the new threshold values to meet the changes as closely as possible.

3.17 Analyzing custom selections of samples from the database

Although samples will often be analyzed per batch, which may correspond to a run on a sequencer, it is also possible to perform analyses on random selections of samples from the database. We refer to the Reference manual, Chapter Database entries for detailed explanation of the comprehensive database querying tools.

17.1 In the HDA sample database, you can simply make a custom selection by selecting a few entries in the *Main* window. This happens by holding down the **Ctrl**-key while clicking on a few entries.

Check boxes for selected entries are indicated as ☑.

17.2 Alternatively, you can select a range by clicking on the first and holding down the **Shift**-key while clicking on the last.

17.3 With a number of entries selected, choose ***Analyze(current selection)*** from the ***HDA*** menu to call the *HDA analysis* dialog box (see Figure 3.36).

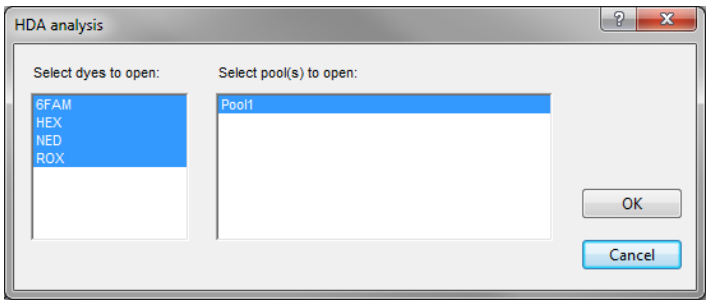

Figure 3.36: The *HDA analysis* dialog box.

The *HDA analysis* dialog box prompts for the pool to analyze and the dyes load.

As opposed to the dialog box that pops up when the *Analysis window* is opened from the *Batch editor* (3.5), the Pool to analyze is required information here. A batch is linked to a set of comparable lanes, which, by nature, can only contain one pool. A sample from the database, in contrast, can have many pools linked to it.

The resulting *Analysis window* and its options are described earlier in 3.5 to 3.12.

This work flow allows you to combine new sample runs with previous runs that can act as reference signals for multiple experiments.

Chapter 4

HDA databasing and reporting functions

4.1 Background information

While importing a batch of HDA profiles into the database, BioNumerics automatically creates *database entries*. These entries are listed in the *Database entries* panel of the *Main* window and initially, they are only characterized by a unique key, which was parsed from the trace name of the .fsa files. In a HDA database setup, the entries correspond to the samples for which HDA profiles have been generated.

More descriptive information fields can be added to the database entries, making it possible to describe the samples adequately and perform searches and queries based on specific database components or combinations thereof.

This chapter gives a brief overview of some basic functions that can be used to start working in a BioNumerics database environment. For more detailed information on the database functions we refer to the Reference manual, Part The BioNumerics database.

4.2 Creating information fields

A number of predefined information fields are automatically created when creating a new database. All predefined information fields are listed in the drop-down menu's in the information fields header of each panel and can be either displayed or hidden. In addition to these default information fields, extra information fields can be added with ***Add information field*** or removed with ***Delete information field***. These menu commands can be accessed in each panel by right-clicking in their information toolbars. Information fields can also be added to the *Database panel* with the corresponding menu commands.

2.1 Click on the *Database entries* panel to highlight it and select ***Edit*** > ***Create new object...*** ().

2.2 Enter the name of the database information field, for example "Origin", and press <***OK***>.

2.3 Select ***Edit*** > ***Create new object...*** () again to define the second field, 'Preparation date'.

2.4 Then, select ***Edit*** > ***Create new object...*** () again to define a third field, 'Sample number'.

2.5 Finally, select ***Edit*** > ***Create new object...*** () again to define a field 'Done by'.

4.3 Entering information fields

3.1 By double-clicking, or pressing **Enter** on a database entry, the *Entry* window appears (see Figure 4.1). Right-clicking on the entry, and selecting ***Open entry*** also works.

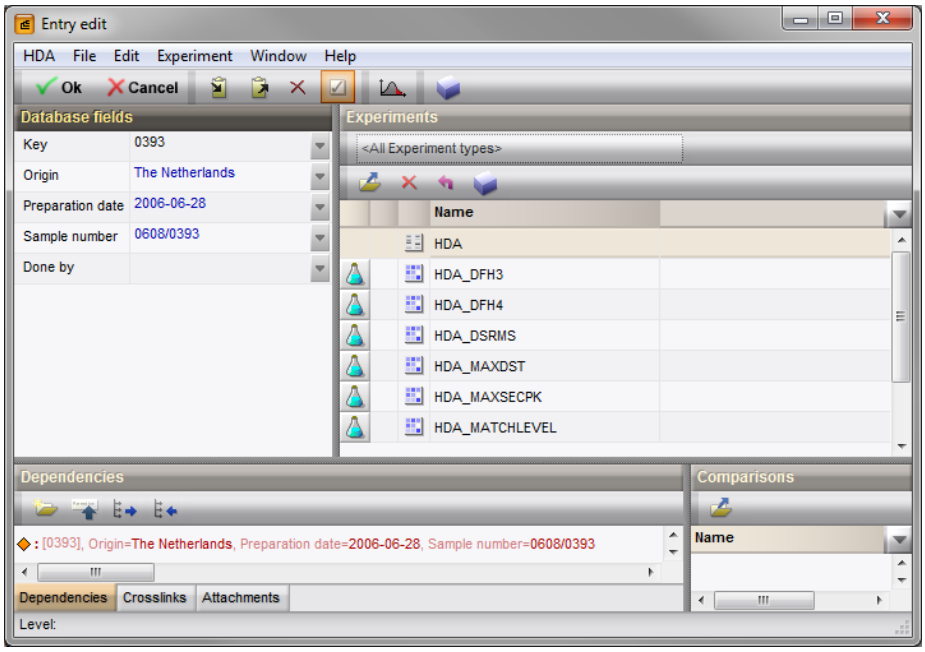

Figure 4.1: The *Entry* window.

3.2 Enter some information in each of the fields (see Figure 4.1).

3.3 If a number of entries have mostly the same fields, you can copy the complete entry information to the clipboard using ***Edit*** > ***Copy to clipboard*** ( ).

3.4 To clear the complete information of the entry, use ***Edit*** > ***Clear all database fields*** ( ).

3.5 To paste the information from the clipboard, use ***Edit*** > ***Paste from clipboard*** ( ).

If some of the information fields are the same as entered for previous entries, you can drop down a history list for each information field. The history lists can contain up to 10 previously entered strings for the information field. Using the history lists is recommended (i) to save time and work and (ii) to avoid typographical errors.

3.6 Drop down a history list by clicking the button on the right hand from the information field. A floating menu appears from which you can select an information string.

3.7 Press the **Enter**-key or <***OK***> to close the *Entry* window and store the information, or press the **Esc** key or <***Cancel***> to close the window without changing any information.

In order to quickly enter the same information for many entries, the use of the keyboard is recommended: use the **ArrowUp** and **ArrowDown** keys to move through the entries in the database, use the **Enter** key to edit an entry, use the **F7** and **F8** keys to copy and paste information, and use the **Enter** key again to close the *Entry* window.

3.8 Alternative to using the *Entry* window, information in non-default information fields in the database (and in other panels containing tabular data) can be edited directly by clicking twice on an information field in the database. The information will appear selected blue against a yellow background and can be edited. If desired, single-click field editing can be enabled via ***File*** > ***Preferences...*** in the *Main* window (see the Reference manual, Chapter The BioNumerics user interface).

4.4 Manual selection functions

As will be explained in 4.6, reporting can be done per batch of profiles, but also from any selection of samples from the database.

A single entry (sample) can be selected by holding the **Ctrl**-key and (left-)clicking. Check boxes for selected entries are indicated as . Selected entries are unselected in the same way.

4.1 Select a single entry (**Ctrl+click**).

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

4.3 If you wish to select entries using the keyboard, you can scroll through the database using the **ArrowUp** and **ArrowDown** keys, and select or unselect entries using the space bar.

4.4 A single entry can be selected or unselected from its *Entry* window (see 4.3) using ***Edit*** > ***Select / unselect this entry*** ( ). When the entry is selected, the corresponding button shows as .

4.5 To show only the selected entries, use ***Edit*** > ***Views*** > ***Switch to Selected Objects view*** (**Ctrl+Shift+S**).

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

4.5 Automatic search and select functions

Besides the manual selection functions as described above, BioNumerics possesses more advanced database search functions.

5.1 Click in *Database entries* panel and select ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**). This pops up the *Find* dialog box.

In the text box next to ***Search for:***, you can type some text that will be searched for in all currently displayed information fields. With every stroke on the keyboard, the display gets updated with those database objects that fulfill the search criteria (= ”hits”). The number of hits is displayed as well (***Found:***). When one or more hits are found, the column ’Hit column’ displays the name of the information field in which the hit was found and the column ’Hit text’ shows the matched string.

Hits are sorted so that the most relevant ones are shown on top of the list:

- Hits that occur at the beginning of an information string are considered more important than hits in the middle of a string.
- Complete matches are ranked higher than partial matches.
- If two search strings are entered (separated with a space), objects that match both searches will appear higher on the list than objects that have a hit for only one of the search strings.

Further typing should narrow the search down to the specific object that you were looking for. When this object is the only hit, it will be automatically be highlighted. If a short list of objects is shown that all fulfill the search criteria, the object can be highlighted by clicking on it. When pressing the <***Highlight***> button or by hitting **Enter** on the keyboard, the highlighted object in the *Find* dialog box will become highlighted in the object grid panel from which the search was launched.

Alternatively, when looking for a group of objects, a selection of objects can be made manually or all objects that fulfill the search criteria can be selected at once by pressing the <***Select all***> button. The objects will also be selected in the object grid panel.

5.2 Press the <***Close***> button to close the *Find* dialog box.

4.6 The HDA match report window

For a batch of HDA traces, or for any selection of samples from the database, a report window can be generated, summarizing the information from all targets used.

6.1 To view a report window for a batch, open the *HDA batch list window* (***HDA*** > ***Edit batches*** in the *Main* window; see also 3.5).

6.2 Select a batch and choose ***HDA*** > ***Report*** to call the *HDA report* dialog box (see Figure 4.2).

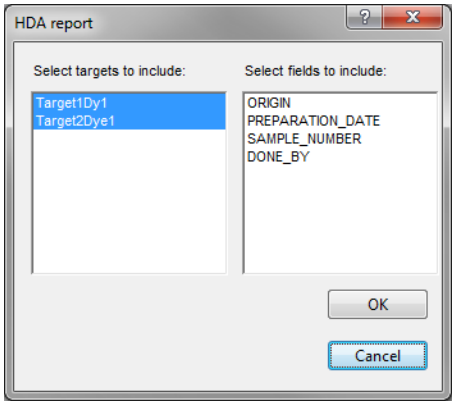

Figure 4.2: The *HDA report* dialog box to select targets and information fields to display in the report window.

The *HDA report* dialog box allows the user to select the targets to use to create the report window as well as the information fields for the samples to show in the report. The keys are always displayed.

The resulting *HDA report window* is shown in Figure 4.3. The left most column shows the sample keys, and immediately right from it, a colored summary of the number of mismatches. As soon as one mismatch is detected for a sample, according to the targets used in the report, the color will be red for that sample, otherwise it will be green. Right from the matching summary, the selected information fields are listed as additional columns. The right most columns are the targets included in the report, indicating the assignment of polymorphic variants or mismatches.

File HDA Window

Table

Sample	Summary	Sample number	Target1Dye1	Target1Dye2	Target1Dye3	Target1Dye4	Target2Dye1	Target2Dye2	Target2Dye3	Target2Dy
0385	2 unmatched	0385-T1	UNMATCHED	WT	WT	WT	WT	UNMATCHED	WT	WT
0386	2 unmatched	0386-T1	UNMATCHED	WT	WT	WT	WT	UNMATCHED	WT	WT
0387	0 unmatched	0387	WT	WT	WT	WT	WT	WT	WT	WT
0388	0 unmatched	0388	WT	FAILED	WT	WT	WT	FAILED	WT	WT
0389	0 unmatched	0389	WT	WT	WT	WT	WT	WT	WT	WT
0390	0 unmatched	0390	WT	WT	WT	WT	WT	WT	WT	WT
0391	4 unmatched	0391-T2	UNMATCHED	UNMATCHED	WT	UNMATCHED	WT	UNMATCHED	WT	WT
0392	2 unmatched	0392-T2	WT	UNMATCHED	WT	WT	WT	UNMATCHED	WT	WT
0393	0 unmatched									

Figure 4.3: The *HDA report window*.

6.3 In case a sample is defined by more than one signal for the same target, the *HDA report window* may display this cell in dark gray. The different results are placed between <> signs and are separated by + signs: <WT+FAILED>. However, the window can behave in two ways, depending on how the samples were selected:

1. If the window was created from a batch (as described in this example), only repeated signals *within the current batch* will be considered. In other words, if a sample within the report window has a repeated signal for the same target but in another batch, this will not be shown.

2. If the window was created from a selection of samples made in the database (see Instruction 5.3), the selection can span multiple batches.

6.4 By clicking in the 'Sample' field, you can select a sample as a whole. By clicking on any of the target cells, the individual cell is selected.

6.5 Click on a target cell and press **Enter** (or ***HDA*** > ***Edit signal***) to call the *Edit HDA signal results* dialog box (see Figure 4.4).

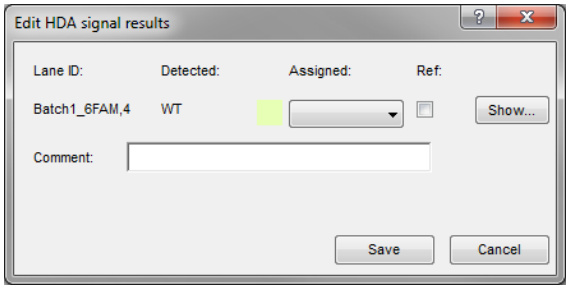

Figure 4.4: The *Edit HDA signal results* dialog box.

The *Edit HDA signal results* dialog box is similar to the *Edit signal* dialog box as described in 3.9. This box, however, does not show the detailed detection/matching information as the *Edit signal* dialog box popped up from the *HDA curves window*.

To display this information, press the <***Show***> button to pop up the detailed signal-to-reference comparison window.

It is possible to manually change the assignment of a signal, to mark a signal as a reference, and to add a comment line.

In case the sample is defined by more than one signal for the selected target, the different signal results may be listed under each other.

6.6 Close the *Edit HDA signal results* dialog box.

6.7 In the *HDA Report window* with a target cell selected, pressing the **Ctrl+Enter** (or ***HDA*** > ***Show signal match***) pops up the *Detailed signal match window* for the selected signal.

6.8 With ***HDA*** > ***Copy to clipboard***, the content of the *HDA report window* is copied to the clipboard as tab-delimited text file. The information can be pasted directly in other programs such as Microsoft Excel.

6.9 Using the print button or with ***File*** > ***Print report***, the report can be sent directly to a printer.

4.7 Creating reports from any selection of samples

By using the selection tools described in 4.4 and 4.5, a selection of samples can be made to create a custom report window.

7.1 With a selection of entries made in the *Main* window, select ***HDA*** $>$ ***Report (current selection)***.

The report window that appears is identical to the one described in 4.6, except that the samples may not be corresponding to one batch.

APPLIED MATHS

A BIOMÉRIEUX COMPANY

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

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

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