BIONUMERICS Tutorial:

Importing and processing AFLP sequencer curve files

1 Aim

Comprehensive tools for processing electrophoresis fingerprints, both from slab gels and capillary sequencers are incorporated into BIONUMERICS. When fingerprints are run on a capillary sequencers the resulting data can have two different formats: curve files (also referred to as electropherograms, chromatogram files or trace files), or peak tables. In this tutorial we will focus on the import of curve files. Since curve files contain the raw data, the fingerprint processing steps (e.g. normalization, band assignment, . . .) are also covered in this tutorial.

2 Sample data

Raw curve files from Applied Biosystems and Beckman sequencers can be imported directly in BIONUMERICS. A batch of Applied Biosystems curve files composing one run can be downloaded from the Applied Maths website (https://www.applied-maths.com/download/sample-data, click on "AB sequencer trace files"). These example files will be used to illustrate the import steps in this tutorial.

3 Importing curve files

1. Create a new database (see tutorial "Creating a new database") or open an existing database.

2. Select *File* > *Import...* (, **Ctrl+I**) to call the *Import* dialog box, choose *Import curves* under *Fingerprint type data* and press <*Import*>.

3. Browse to the downloaded and unzipped example data folder and select all Applied BioSystems curves files (extension `.fsa`). Press <*Open*>.

The files are displayed in the *Input* wizard page and the default suggested *Fingerprint file name* is the folder name.

4. Change the name to **Run1** and press <*Next*> (see Figure 1).

The way the information should be imported in the database can be specified with an import template. In the example data set, the dye name can be parsed from the file information and the
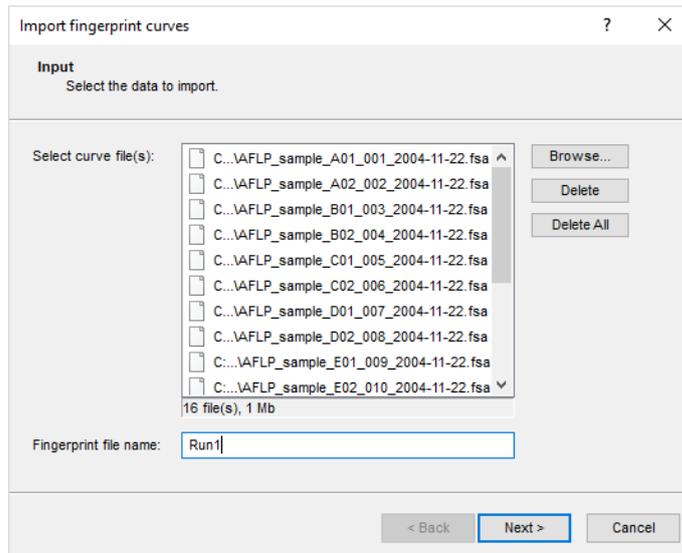
Figure 1: Selected curve files and fingerprint file name.

unique sample information can be parsed from the file names. A new import template needs to be defined:

    5. Press the **Create new** button to call the *Import rules* dialog box (see Figure 4).

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

    6. Select **Curve dye** from the list, select <**Edit destination**> and select **Fingerprint dye** as corresponding field. Press <**OK**>.

Next, we will specify a new rule that links the part of the file name appearing after the third and before the fourth underscore to the **Key** field.

    7. Select **Name** from the list, select <**Edit destination**> and select **Key** as corresponding field. Press <**OK**>.

    8. Check the option **Show advanced options**, make sure the last row is selected in the grid panel and press the <**Edit parsing**> button.

    9. In the *Data parsing* dialog box, fill in following data parsing string: "*_*_*_[DATA]_*". The asterisk will serve as wildcard.

    10. Press the <**Preview**> button and press <**OK**> when the parsing is correct (see Figure 2).

In the example files, "Lane" information is available. We will import this information and store it as fingerprint information.

    11. Select **Lane** from the list and select <**Edit destination**>. Select **Create new** under **Lane info field** and press <**OK**> twice and confirm the creation of the new field.

The *Import rules* dialog box should now look like Figure 4.

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

In the example data set, the ROX channel contains the size standard (GeneScan 500 ROX) and the 5-FAM channel represents the actual samples.

    13. Make sure **ROX** is selected as reference dye and uncheck the dyes **JOE** and **NED** to prevent the import of these channels (see Figure 5).
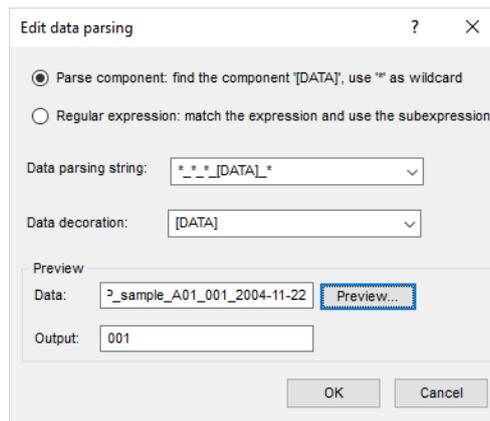
Figure 2: Data parsing string.
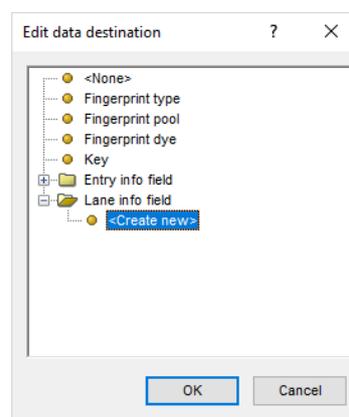
Figure 3: Link to lane information field.

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

15. Specify a template name, e.g. **Import AB curve files** and press <***OK***>.

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

The preview should now look like Figure 6.

17. Close the preview.

18. Make sure ***Create new*** is selected as fingerprint type experiment, select the new template and press <***Next***>.

19. Specify a name for the new base fingerprint type experiment (eg **AFLP**) and press <***OK***>. Confirm the creation of the new experiment in the database.

20. Confirm the proposed OD range.

A fingerprint type needs to be present in the database for each pool (if present) and dye combination. The names of these fingerprint types are composed of the base fingerprint type name, followed by the pool name (if present), and the name of the dye (e.g. AFLPROX). If one or more of these fingerprint types are not present in the database, a new dialog box pops up, listing all missing fingerprint types (see Figure 8).

21. Confirm the creation of the two missing fingerprint type experiments.
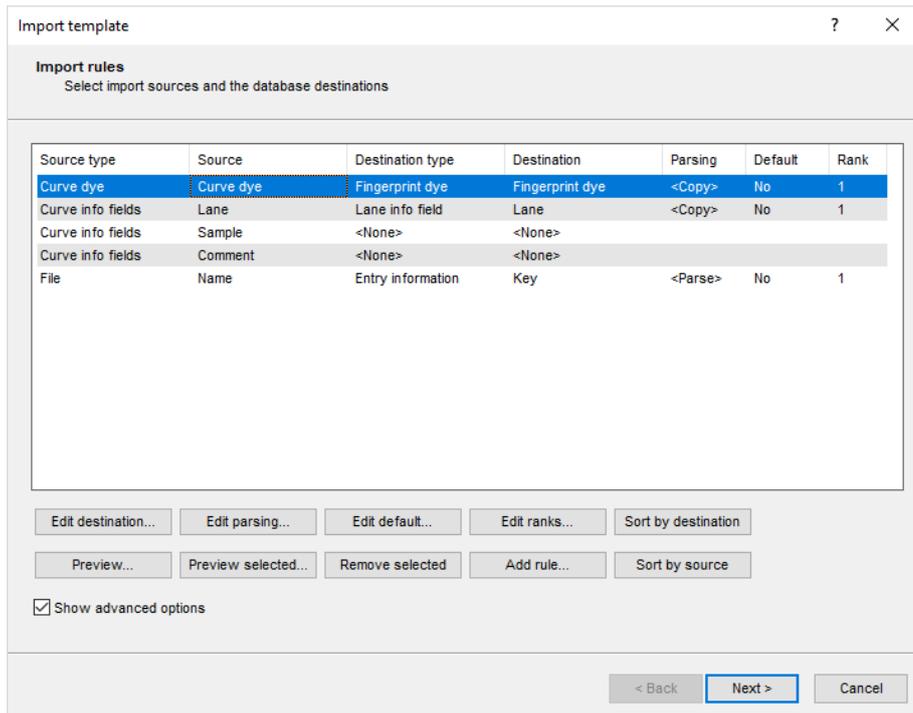
Figure 4: Import rules.
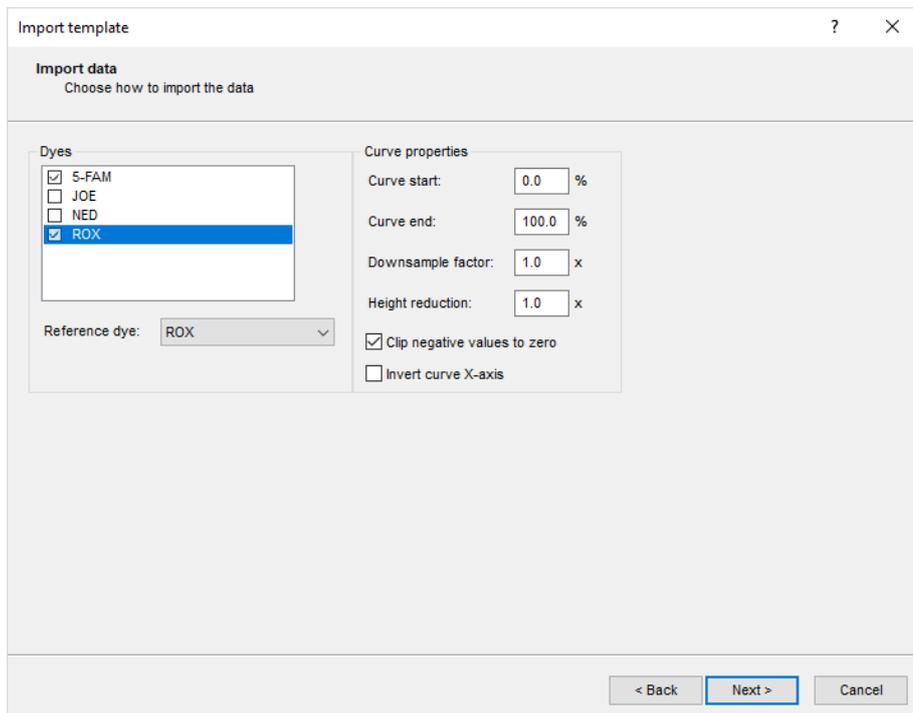
Figure 5: Specify how to import the data.

22. Press <***Next***> to confirm the creation of new entries (see Figure 9).

23. Make sure ***Open curve preprocessing window*** is checked in the last step and press <***Finish***>.

For each dye checked in the *Dyes panel* of the *Import data* dialog box, a new fingerprint file is created, composed of the file name specified and the name of the dye (e.g. Run1_ROX). These

Figure 6: Preview of the parsing.
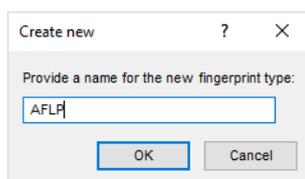
Figure 7: New base fingerprint type experiment.
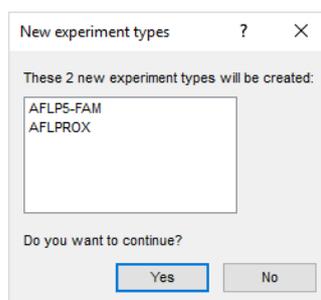
Figure 8: Confirm creation of experiments.

files are displayed in the *Fingerprint files* panel. The reference file is shown in the **Link** column. Double-clicking on a fingerprint file opens the *Fingerprint* window. If lane information was imported with the individual lanes, this information is displayed in the *Fingerprint information panel*.

The imported fingerprint lanes are linked to new entries in the database. The lanes are linked to the corresponding fingerprint "dye" type. The names of these fingerprint types are composed of the base fingerprint type name, followed by the pool name (if present), and the name of the dye. The fingerprint type experiments are displayed in the *Experiment types* panel.

After data import, the *Main* window looks as in Figure 10.
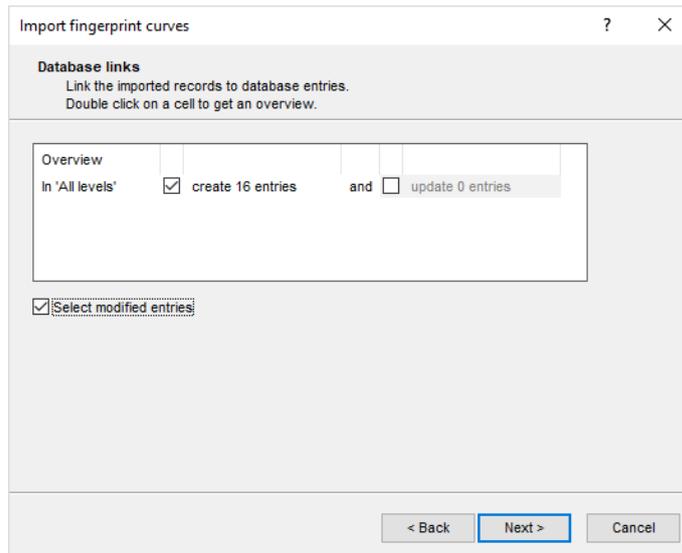
Figure 9: Database links.
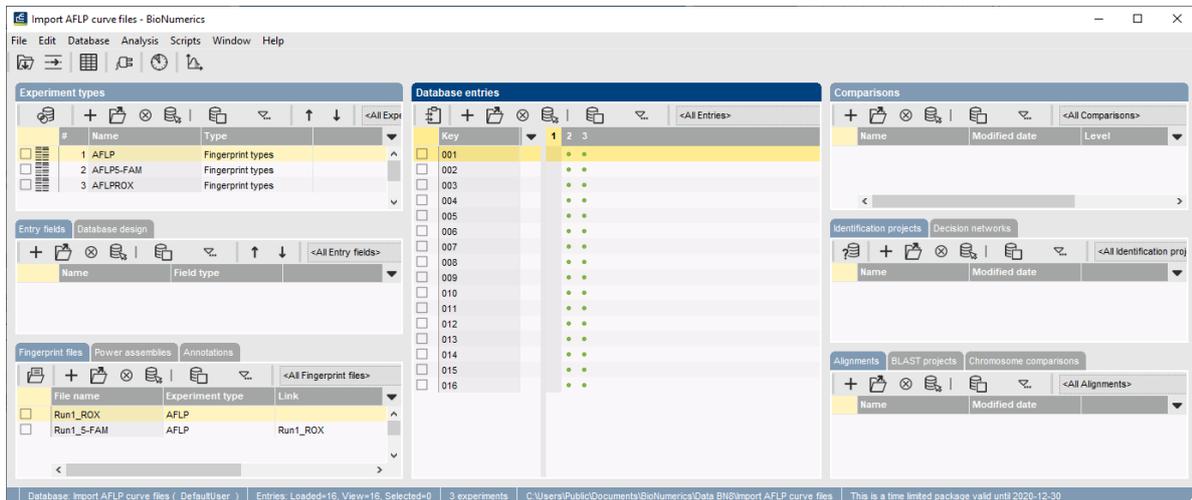
Figure 10: The *Main* window after import of the data.

4   Processing curve files

When the option ***Open curve preprocessing window*** was checked in the last step of the import routine, the *Fingerprint curve processing* window opens when pressing the <***Finish***> button. The two channels from the run are automatically loaded and displayed in the *Fingerprint curve processing* window.

> The *Fingerprint curve processing* window can also be called from the *Main* window by highlightig one of the channels in the *Fingerprint files* panel and select ***Open finger-print data...*** (🖻). Alternatively, you can first open the *Fingerprint* window with ***Edit*** > ***Open highlighted object...*** (🗁, **Enter**) and then select ***File*** > ***Edit fingerprint data...*** (🖩).

1. Select ***File*** > ***Curve processing settings...*** (〽), change the ***OD range*** to "4096" (= 12-bit) in the *Curve processing settings* dialog box and press <***OK***>.

2. Click on the 👁 icon left of the **Run1_5-FAM** channel in the *Channels* panel.

The data channel is now hidden from the view and its icon is displayed as 👁.

> 3. Use the zoom sliders on the left and on top to optimize the display of the fingerprint curves.

Since the raw chromatogram files have not undergone any preprocessing, normalization will have to be performed. This requires a *reference system* to be defined, based upon the marker peaks available in the reference dye.

> 4. Make sure the reference dye is the only dye visible in the upper panel.
>
> 5. Select **Bands** > **Search reference bands...** (📷, **Ctrl+F**) to call the *Search reference bands* dialog box.
>
> 6. Specify a peak detection **OD range** of **2** (in %) and a peak detection **Curve range** of **1** (in %). Press <**OK**>.

The bands that fall within the specified criteria are marked with a solid line at the band's position (see Figure 11).
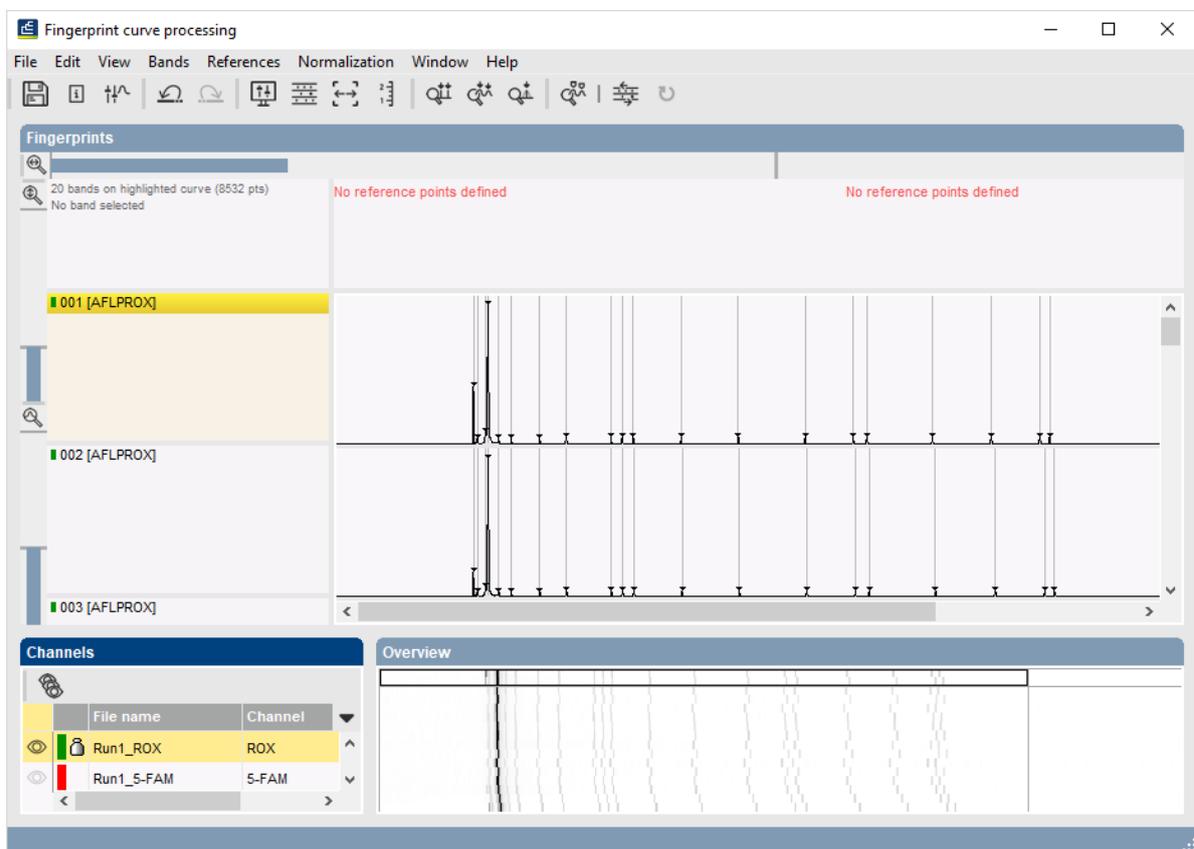
Figure 11: The *Fingerprint curve processing* window only displaying the reference dye.

> 7. To have a reference system automatically created based on a lane containing commercial size marker, first highlight a suitable lane and then select **References** > **Define size standard...**.

This will display the *Size standard* dialog box, from which a size marker can be selected (see Figure 12).

In the example curve files, the GeneScan 500 ROX size standard is used as reference, containing 16 bands with known molecular weight.

> 8. Select **GeneScan 500 ROX** from the list, select **Pattern match** and press <**OK**>.
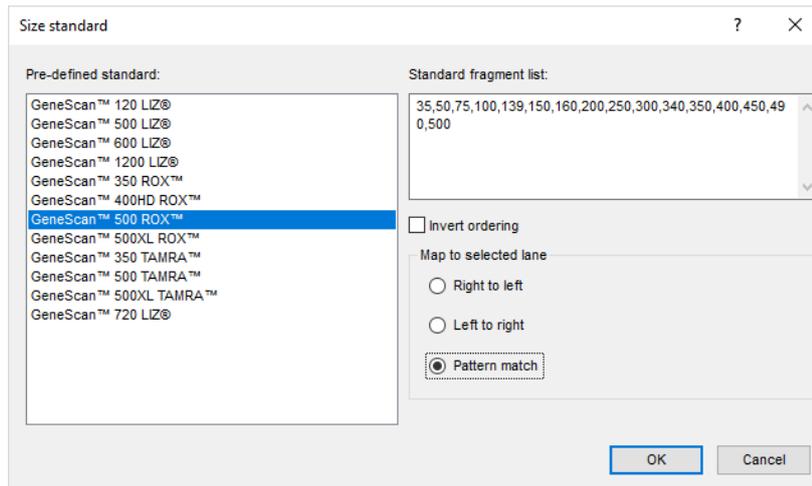
Figure 12: Choose a size standard from the list.

9. Save the data to the database with **File** > **Save** (🖫, **Ctrl+S**).

The software will automatically create the reference system and calibration curve for each of the fingerprint types. Since this allows the calculation of metrics information, a metrics scale now becomes available in the upper part of the *Fingerprints* panel.

Normalization is achieved by assigning bands in the reference channel to external reference positions.

10. To normalize a complete run at once, select **Normalization** > **Auto assign reference positions (all lanes)...** (⬚, **Ctrl+A**), leave all settings unaltered and press <**OK**>.

11. When the assignment of the marker bands to reference positions is made, the data can be shown in normalized mode with **Normalization** > **Show normalized view** (⬚, **Shift+N**).

12. Click on the 👁 icon left of the **ROX** and **5-FAM** channels in the *Channels* panel.

The data channel is now shown and the reference channel is hidden from the view.

13. Select **Bands** > **Search data bands...** (⬚, **Ctrl+Shift+F**) to call the *Search data bands* dialog box.

14. Specify a peak detection **OD range** of **2** (in %) and a peak detection **Curve range** of **1** (in %). Check **Filter by fragment length**, specify a **Minimum fragment length** of **35** and press <**OK**>.

15. Optionally the *Search data bands* dialog box can be called again to optimize the band search settings.

✏ If you want to calculate the similarity between the AFLP data curves in the *Comparison* window using a curve based coefficient (e.g. **Pearson correlation**), the search for bands can be skipped in the data channel.

16. Save the changes and close the *Fingerprint curve processing* window.

5 Conclusion

In this tutorial you have seen how to import and process sequencer curve files in BIONUMERICS. You can now start applying comparison functions such as band matching, clustering, etc. on the
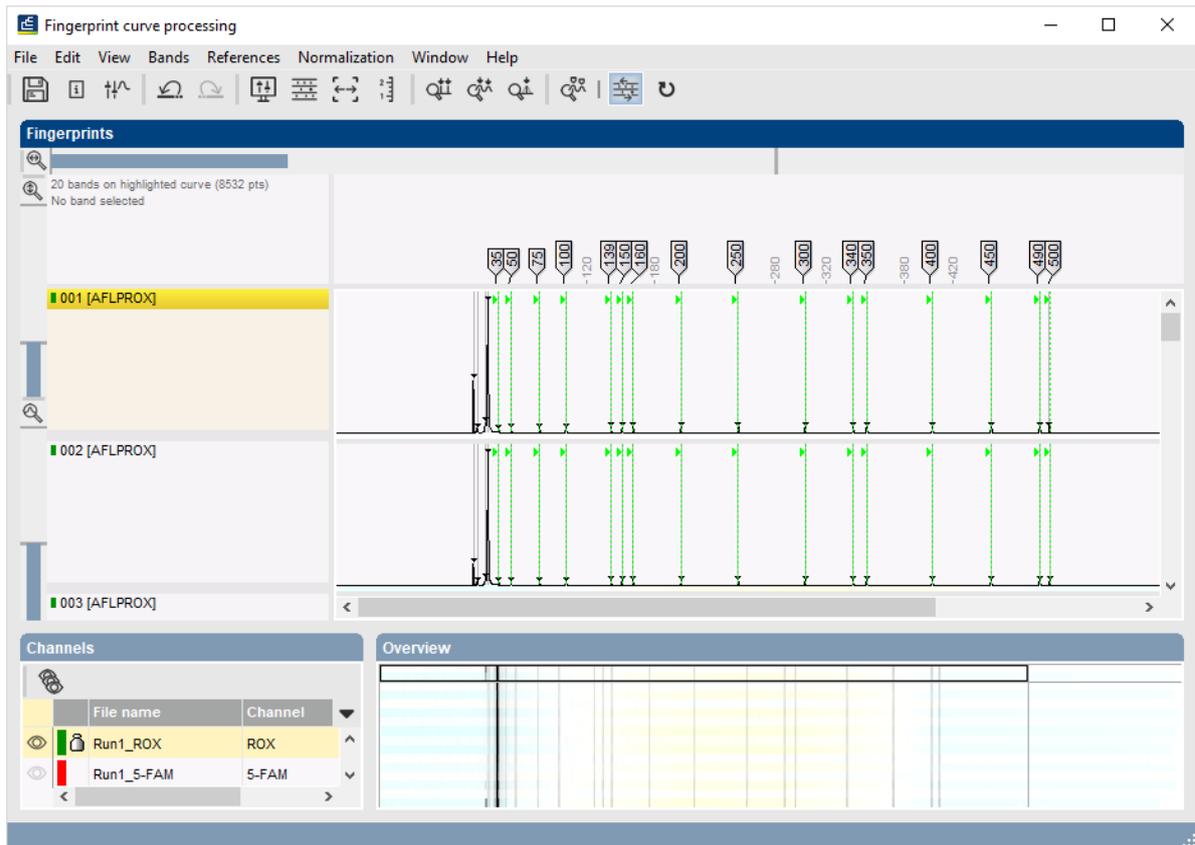
Figure 13: Normalized view - reference channel.

data. More information about these functions can found in the analysis tutorials on our website.