User Manuel for RisaAligner software

Objective

One of the most popular fingerprinting methods to compare microbial communities is the technique called Ribosomal Intergenic Spacer Analysis (RISA). RISA profiles can be generated from PCR products with the Agilent 2100 Bioanalyzer using DNA 1000 or High Sensitivity chips. The 2100 expert software plots fluorescence intensity versus size/migration time and produces an electrophoregram for each sample. The Agilent Expert software generates a peak table for each sample aligned on the ladder theoretical sizes. These data can be exported and further analyzed. For many applications, working with peak tables is sufficient, but for community analysis more information can be obtained with electrophoregram data. However repeatability among chips cannot always be obtained because of variations in ladder migration from one chip to the other. This is a major issue when several electrophoregrams have to be analyzed jointly. In order to resolve this problem we propose a realignment algorithm which corrects the data in such a way that all the ladders match, and therefore the data obtained from different chips can be meaningfully compared.

Detailed procedure

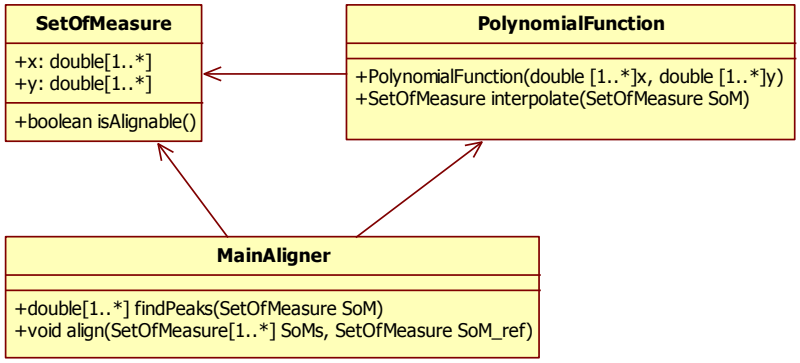
Exporting data from the Agilent Bioanalyzer

After performing the baseline correction with the Bioanalyzer Expert Software to obtain a flat baseline for all samples (select: Global, Advanced mode and in Ladder Setpoints select baseline correction), data must be exported (in File, Export, select: Sample Data, aligned sample data) in a dedicated folder (creates csv files, one file per sample and the ladder). If not exported from the Bioanalyzer Expert software, csv files must be generated with the following format:

```
22-23-24-25-26-27-28-29-30-31_2013-02-21_16-29-57.xad
C:\Users\A\Desktop\these\profils Risa
jeudi 21 fevrier 2013 15:29:57
vendredi 22 fevrier 2013 09:02:50
B.02.07.SI532
B.02.07.SI532

High Sensitivity DNA
C:\Program Files\Agilent\2100 bioanalyzer\2100 expert\assays\dsDNA
High Sensitivity DNA Assay
1.0

                                                    11

Ladder
                                                    2121

Value
9.65252E-06
0.
0.
0.
0.
0.
0.
0.
1.427703E-04
2.447061E-04
2.812718E-04
3.813987E-04
2.387612E-04
1.638977E-04
2.125681E-04
1.777638E-04
7.166899E-05
2.98618E-05
.
.
.
.
```

Programming

The software is programmed in JAVA to ensure a maximum compatibility with all kind of computer. The user interface has been implemented by using both existing and home-made classes, the description of which is out of the scope of this user manual. The core of the software is implemented by the classes, as depicted in the simplified UML class diagram hereafter:



The class **SetOfMeasure** represents a dataset. The method **isAlignable** returns the value true if the dataset can undergo the primary alignment. The class **PolynomialFunction** defines an interpolation polynomial, given a set of interpolation points (**x**,**y**). In the present implementation, cubic spline interpolation has been used. Given an instance of **SetOfMeasure**, the method **interpolate** returns a new instance of **SetOfMeasure** where the data have been modified accordingly to the defined interpolation points.

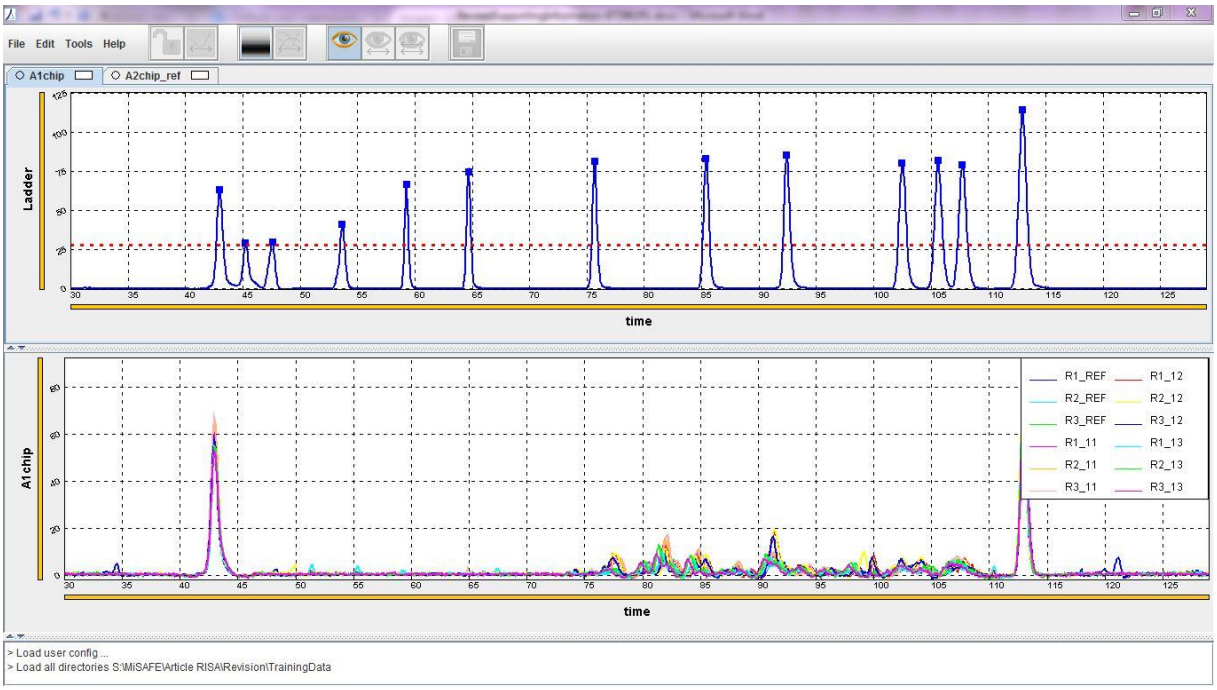
The main class **MainAligner** contains the methods for performing the primary and secondary alignment. In particular, the method **findPeaks** returns a chip of coordinates which correspond to the peaks which have been detected in a given dataset (i.e. in instance of **SetOfMeasure**). The method **align** performs the alignment, given a chip of datasets (**SoMs**) and a reference dataset (**SoM_ref**). A JAVA-like pseudocode of the core of this method is reported hereafter:

```
public void align(SetOfMeasure[] tab_SetOfMeasure, SetOfMeasure setOfMeasure_ref)
{
   double[] ref_peaks_x = findPeaks(setOfMeasure_ref);

   // loop on all opened set of measures
   for (int i=0 ; i<tab_SetOfMeasure.length ; i++) {
      SetOfMeasure setOfMeasure = tab_SetOfMeasure[i];
      if (setOfMeasure.isAlignable()) {
         // get x values of peaks of current curve
         double[] peaks_x = findPeaks(setOfMeasure);

         // computation of Polynomial Spline Function to align x values of peaks
         // from current curve on reference curve
         PolynomialFunction psf1 = PolynomialFunction(peaks_x, ref_peaks_x);

         // transformation of current curve with this polynomial Spline Function
         SetOfMeasure setOfMeasure_tmp = psf1.interpolate(setOfMeasure);

         // computation of Polynomial Spline Function for this new curve
         PolynomialFunction psf2 = PolynomialFunction(setOfMeasure_tmp.x,
                                                      setOfMeasure_tmp.y);

         // computation of same curve with original x_values
         setOfMeasure = psf2.interpolate(setOfMeasure_tmp);
      }
   }
}
```

Opening files

Open the software by a double click on the RisaAligner.jar executable file. Each dataset can be imported from a folder containing all **.csv** files by using the menu *File – Open chip*. It is also possible to import several datasets with a single action by the menu *File – Open several chips (Example: TrainingData)*. In this case, the user will be asked to select a directory containing all the folders with the csv files. These directory and folders must exclusively contain chip data.
For each dataset, a tab is opened in the upper side of the main window (Figure S1).

Figure S1: RisaAligner Dashboard.



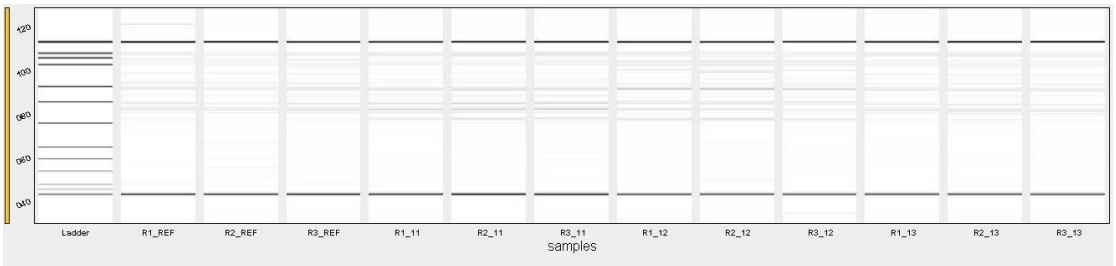
All the operations performed during a study can be saved into a project folder, and at any time it is possible to save the present state of the study, or restore it from an existing project. A project can be saved through the menu *File – Save project*. The menu *File – Open project* allows restoring a study from a project file. The extension of the project files is **.jof**.

During a study, all operations are automatically recorded in a project, the name of which is **#saveFile.jof**. In the case of a sudden shutdown of the software, the user will be asked if he wants to restore the file to the state before the shutdown.

Data display

In the upper window, the ladder is presented (Figure S1) and in the lower window, a single dataset (selected through the tab) is displayed either as a curve Figure S1), or as a gel image (Figure S2).

Figure S2. Gel image of the selected DNA 1000 chip data visualized with the RisaAligner.

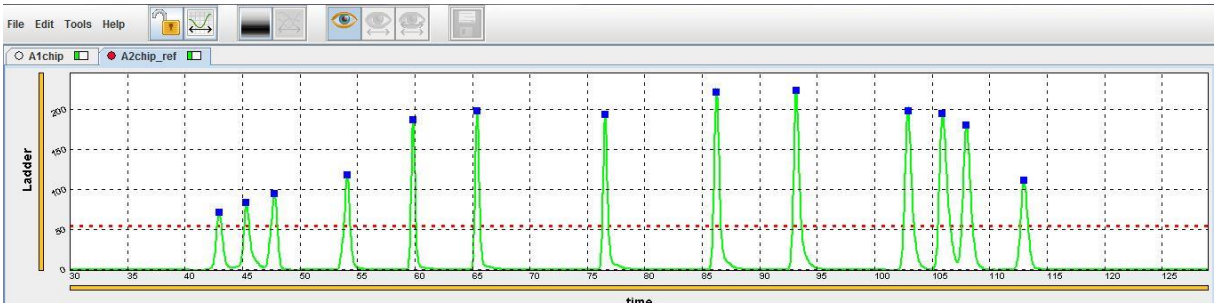


The user can switch between these two graphics with the button  . These graphics can be zoomed with the mouse wheel, moved (mouse left button) or restored (mouse right button).

Selection of the reference ladder

The first step of alignment consists of selecting a reference ladder by clicking on the round button in the left side of the tab. The reference tab will be identified by a red spot (in the training data, the reference is "A2chip_ref") and the ladder electophoregram (Figure S3) will be draw in green (previously blue).

Figure S3. Electrophoregram displaying the selected reference ladder.



Alignment of the other ladders

The software aims at aligning each ladder with the reference. The alignment map is computed based on some characteristic points in the ladders. The number of these characteristic points must be identical for all the ladders.

The software automatically identifies ladder peaks that will be used for alignment (detected peaks have a blue square on the top). A horizontal dotted red line defines the threshold limit and can be moved to detect more peaks. Moreover, it is possible to enable (green square) or exclude (red square) each peak independently by clicking with the left mouse button on the squares (Figure S4).

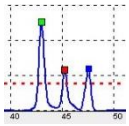
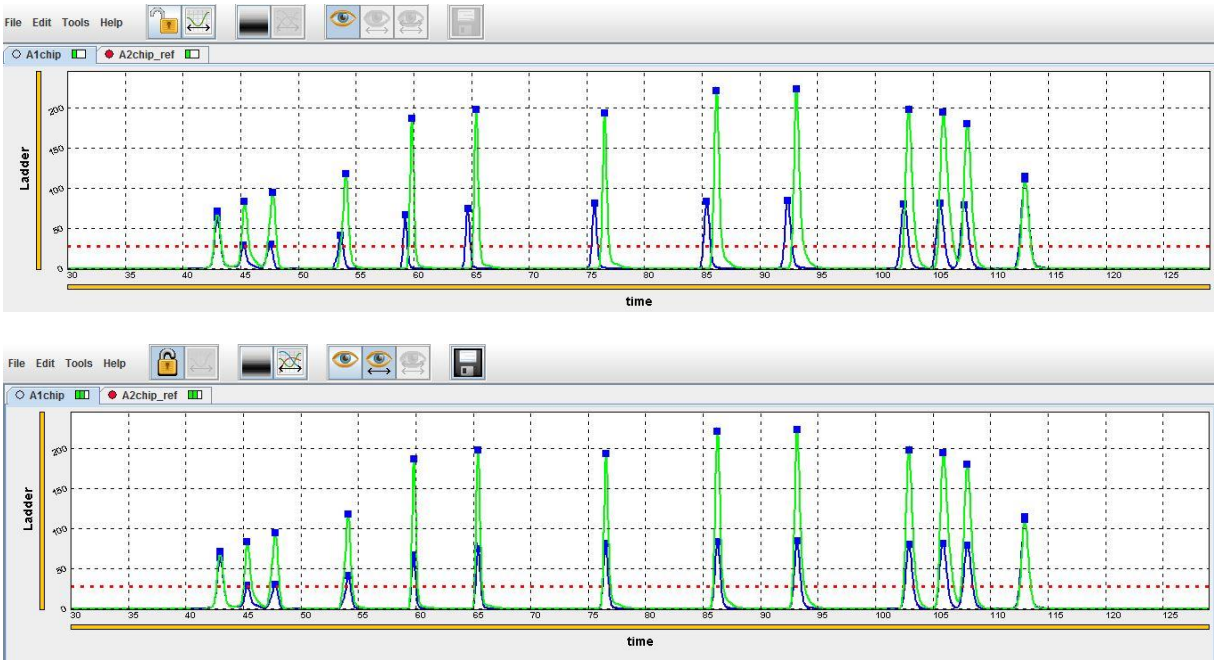


Figure S4: Electrophoregram displaying the peak state. Blue Square: automatically detected peak, Green Square: manually enabled peak, Red Square: manually excluded peak, red dotted line: threshold level.

As soon as the number of enabled peaks is identical to the reference, a ladder becomes align-able, and a green bar appears in its tab (Figure S3). When all the ladders are align-able, the user can start the algorithm by using the button  or the menu *Align ladder*.

The modified (i.e. aligned) ladders are marked by two green bars in their respective tabs (Figure S5). Aligned dataset are locked  (the lock button shift from open to close on the left top side of the tab) and it is not possible to select a different reference nor change the characteristic points used for the alignment except when undo is performed by clicking on the lock button.

Figure S5: Alignment of ladders. Upper image: before alignment, Lower image: after alignment; In Green: the reference ladder, in Blue: aligned ladder of the selected chip (A1 chip), green tabs: the presence of two green tabs means that alignment was performed, lock button: no modification allowed when closed.



Exporting the aligned dataset

The aligned datasets can be exported as **.csv** files by using the button  or the menu *Export*. The data are normalized between 0 and 100 before exporting. By default, datasets are exported in a new directory named **aligned_datas** inside the directory of the project.

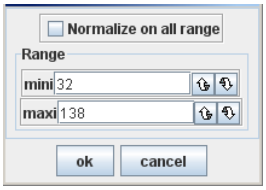
Warning: the original files generated by the Bioanalyzer may contain negative values. The software may introduce some other negative values during the interpolation process. Negative values have no biological meaning. Therefore, each dataset is corrected according to two procedures:

- one with the extension **_aligned.csv** where all negative values were replaced by 0,
- one with the extension **_aligned_negat.csv** where the horizontal axis was shifted to the lowest negative value, so the minimum value became 0.

During the export process, the user is asked through a dialog window (Figure S6):

- if the normalization must be performed on the whole dataset or only on a selected range. The range must be at least restricted from 45 to 110s to suppress the first and the last peaks corresponding to the Upper and Lower Markers added to all samples during the chip preparation.

Figure S6. Dialog box for export procedure

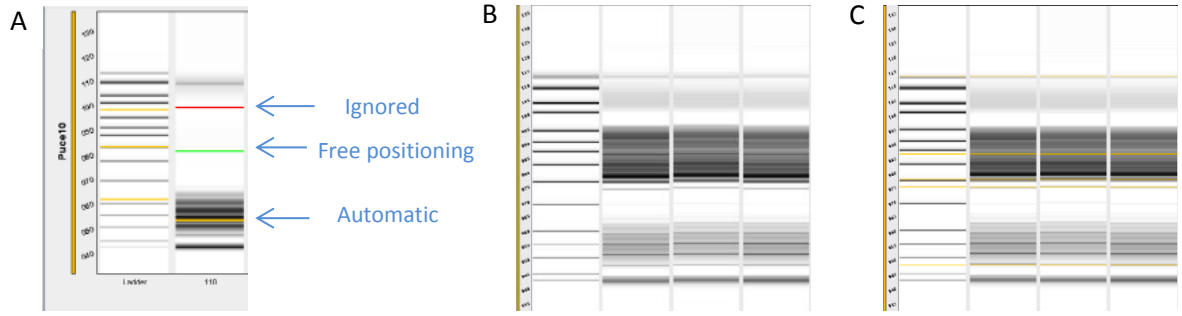


Finally, for each chip, two files per sample and the ladder are generated with the extension _aligned.csv and _aligned_negat.csv and two synthetic files containing all the samples and the ladder with the chip name followed by _aligned.csv and _aligned_negat.csv.

Secondary alignment

The secondary alignment can be performed only with the gel image representation. For each sample, the user can click in the ladder profile in order to add characteristic points which will be used to align the bands within a chip. For each sample, these characteristic points (which must be manually identified by an experienced user) can be modified by using the mouse. Their attribute can be modified by clicking (left mouse button) on it so as to switch between the three states (Figure S7A).

Figure S7. Second alignments with gel reconstituted image. (A) Yellow line (automatic option): the characteristic point will be automatically placed on the closest band; green line (free positioning option): the characteristic point can be set anywhere – i.e. the software will let the user place it without any constraint, red line: the characteristic point will not be taken into account for the alignment. (B) First step aligned data. (C) Second step aligned data.



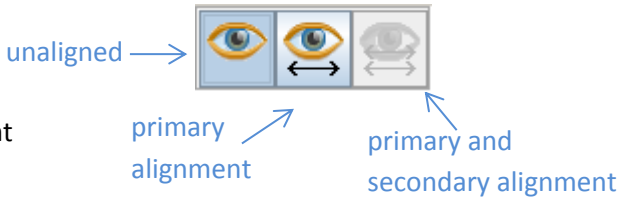
Some samples have identical profiles but with a shift (Figure S7B). We recommend selecting at least two characteristic points (the upper marker and one of the shifted bands) to perform the secondary alignment. All characteristic points must be accurately positioned in each sample by zooming on the profiles. When positioning is perfectly performed, the secondary alignment can be run by pressing the button  or by the menu *Align sample* (Figure S7C). The dataset for which the secondary alignment has been performed are identified by three green bars on their own tab (that is: 0 bar = unaligned, 1 bar = align-able, 2 bars = primary alignment performed, 3 bars = primary and secondary alignments performed).

As for the primary alignment, the aligned datasets can be exported as **.csv** files by using the button  or the menu *Export*. If the secondary alignment has been performed, two additional sets of files with the extensions **_aligned2.csv** and **_aligned2_negat.csv** are generated.

Display tools

For each dataset, it is possible to switch between the three display modes through the following buttons:

- unaligned (i.e. raw data)
- primary alignment
- primary and secondary alignment



Printing

It is possible to print the graphs by the menu *File – print*. Users can choose to print the ladder window or the sample window either as electrophoregram or gel image depending on the selected option.

Citation

When publishing results aligned with this software, please cite the paper “RisaAligner software to align fluorescent data between Agilent 2100 Bioanalyzer chips; application to soil microbial community analysis”. BioTechniques BT5962R1, 2015 by:

Elisabeth Navarro[1-2], Olivier Fabrègue[1], Riccardo Scorretti[1], Jérémy Reboulet[1], Pascal Simonet[1], Lorna Dawson[3] and Sandrine Demanèche[1*.]

1 Université de Lyon, Laboratoire Ampère (CNRS UMR5005), Environmental Microbial Genomics, École Centrale de Lyon, Ecully, France

2 IRD-UMR LSTM, Campus de Baillarguet, 34398 Montpellier, France

3 James Hutton Institute, Aberdeen, UK

* Corresponding author: Sandrine DEMANECHE, sandrine.demaneche@ec-lyon.fr