STATFINGERPRINTS

Processing and statistical analysis of molecular fingerprint profiles

Rory J. Michelland and Laurent Cauquil

Contact: StatFingerprints@gmail.com

Welcome to STATFINGERPRINTS.

This program is a free package for the R statistical program with a user-friendly graphical user interface (GUI). It is intended to help microbial ecologists to analyse fingerprint profiles. It provides procedures both to process and to perform numerous univariate and multivariate statistical analyses on fingerprint profiles. STATFINGERPRINTS is also able to plot fingerprint profiles and several graphical results in 2 or 3 dimensions. It supports import and export of all ASCII files, a format easily writable or readable with text editors, plus the ability to convert FSA files from an ABI Prism sequencer into ASCII files. For advanced use, all procedures can be executed from the R prompt.

The processing part contains procedures:
- to align fingerprint profiles;
- to delete background under peaks and carry out parameterization;
- to homogenise the baseline between fingerprint profiles;
- to normalise fingerprint profiles with 3 algorithms;
- to transform into presence/absence the raw fingerprint profiles with parameterization possibilities for peak detection;
- to estimate 6 ecological diversity indexes from fingerprint profiles and parameterize peak detection;
- to correct defective peaks.

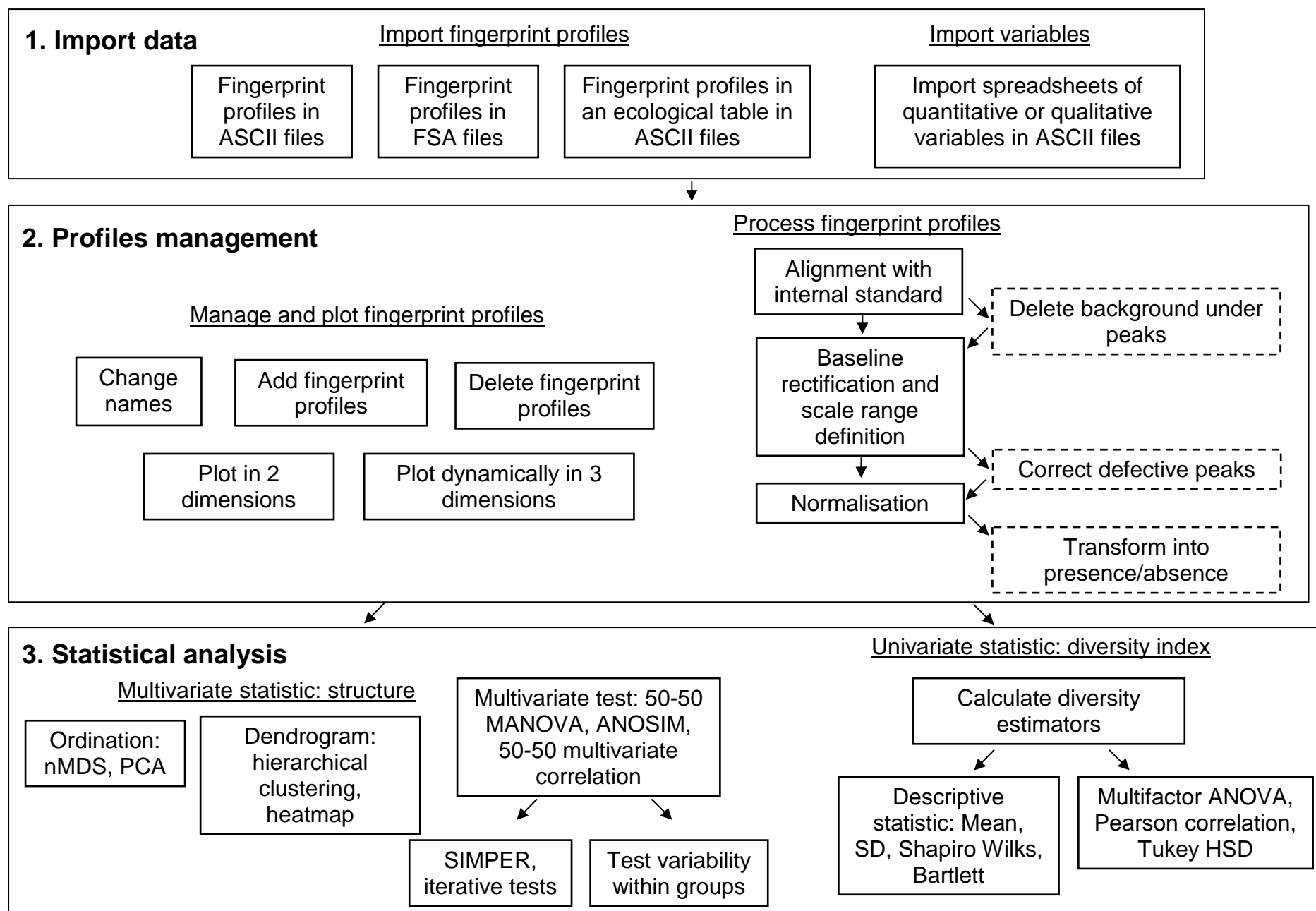The multivariate statistical part contains the following possibilities:
- random start non-metric MultiDimensional Scaling (nMDS) with 2- or 3-dimensional dynamic plotting and the ability to add qualitative and quantitative variables;
- Principal Components Analysis (PCA) with 2- or 3- dimensional dynamic plotting and the ability to add qualitative and quantitative variables;
- Hierarchical clustering with 13 similarity measures and 7 plotting algorithms;
- Heatmap with 13 similarity measures;
- 50-50 Multivariate ANalysis Of VAriance (50-50 MANOVA);
- Analysis Of SIMilarity with the global and pairwise algorithms;
- Analysis of within-group variability;
- SIMilarity PERcentages Procedure (SIMPER);
- Iterative tests (T test, Mann Whitney and Fisher's exact test) to define areas presenting differences along fingerprint profiles;
- 50-50 multivariate correlation.

The univariate statistical part contains the following possibilities:
- Shapiro Wilks test;
- Bartlett test;
- Multifactor ANOVA;
- Pearson correlation.

I and my colleagues will be happy to help you if you encounter any problems (StatFingerprints@gmail.com).

Fig 1: Organigram of the program. Arrows indicated consecutive procedures.

1. Import data

Import fingerprint profiles

Fingerprint profiles in ASCII files	Fingerprint profiles in FSA files	Fingerprint profiles in an ecological table in ASCII files

Import variables

Import spreadsheets of quantitative or qualitative variables in ASCII files

2. Profiles management

Process fingerprint profiles

Manage and plot fingerprint profiles

Change names	Add fingerprint profiles	Delete fingerprint profiles

Plot in 2 dimensions	Plot dynamically in 3 dimensions

Alignment with internal standard → Delete background under peaks

Baseline rectification and scale range definition → Correct defective peaks

Normalisation → Transform into presence/absence

3. Statistical analysis

Multivariate statistic: structure

Ordination: nMDS, PCA

Dendrogram: hierarchical clustering, heatmap

Multivariate test: 50-50 MANOVA, ANOSIM, 50-50 multivariate correlation

SIMPER, iterative tests

Test variability within groups

Univariate statistic: diversity index

Calculate diversity estimators

Descriptive statistic: Mean, SD, Shapiro Wilks, Bartlett

Multifactor ANOVA, Pearson correlation, Tukey HSD

3

CONTENTS

1- INTRODUCTION TO STATFINGERPRINTS

1-1 Installation and running

The STATFINGERPRINTS program and R work on a wide variety of UNIX, Windows and MacOS platforms. Before installing STATFINGERPRINTS you first need to install R (R development Core Team, 2008). Sources, binaries and documentation for R can be obtained via CRAN, the "Comprehensive R Archive Network" (http://cran.r-project.org/mirrors.html). To install the STATFINGERPRINTS program, write at the R prompt:

> install.packages("StatFingerprints",dependencies=T)

Then select the CRAN mirror and the STATFINGERPRINTS program will be downloaded and installed.

Once you start R, you don't need anymore to re-install the STATFINGERPRINTS program but you need to load it by writing at the prompt:
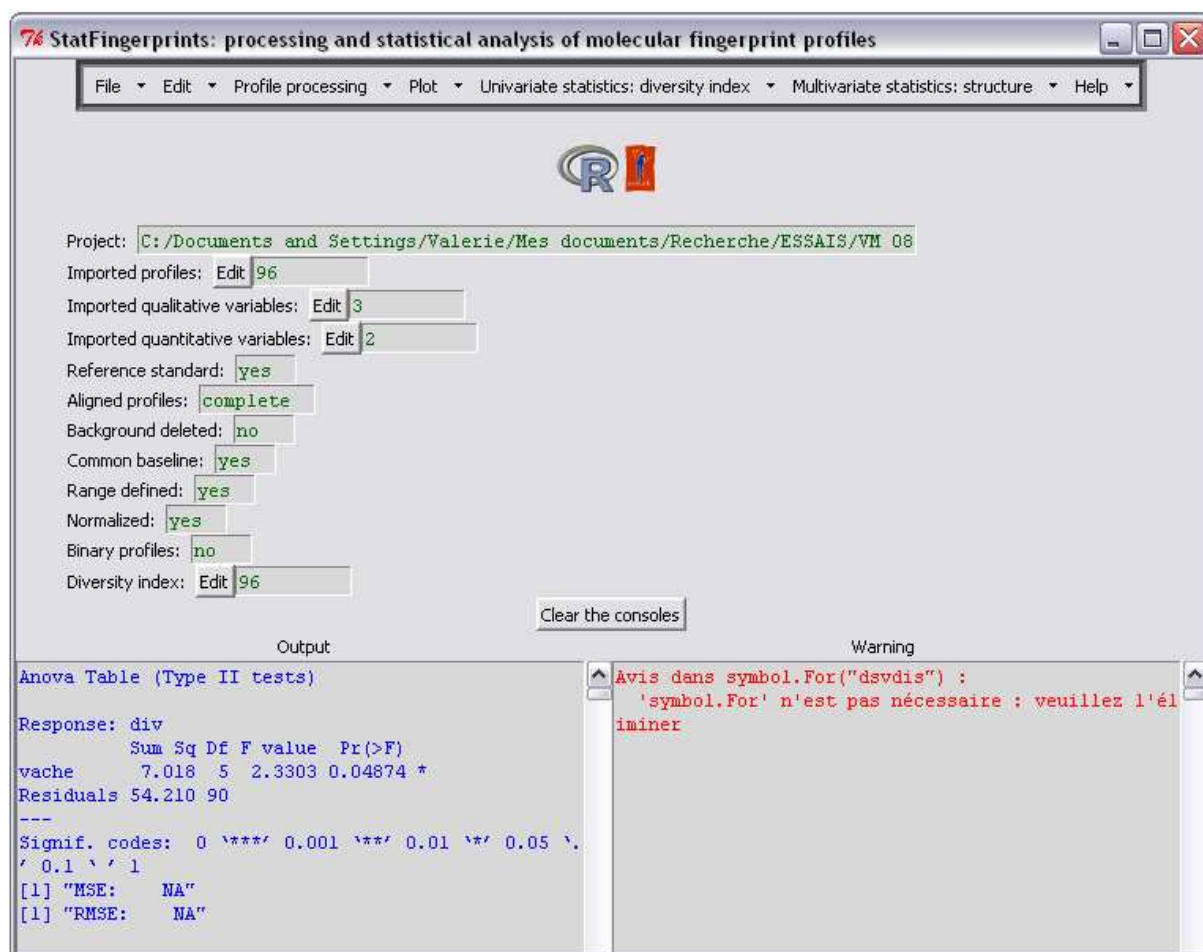
> library(StatFingerprints)

Once loaded, you can write at the prompt first one of the following line command to run STATFINGERPRINTS:

> StatFingerprints()
or
> StatFingerprint()
or
> statfingerprints()
or
> statfingerprint()
or
> SF()
or
> sf()

1-2 Presentation of the main window

Fig 2: The main window of the program.

The main window of the STATFINGERPRINTS program is divided into three parts (Fig 2). In the menu bar on the top, all procedures of the program can be launched using one of the 7 menus. In the status part on the bottom, the number of the fingerprint profiles, those of the qualitative variables, those of the quantitative variables can be seen. The name of the fingerprint profiles, those of the qualitative variables, those of the quantitative variables and those of diversity index can be changed by using the EDIT button (Fig 3).
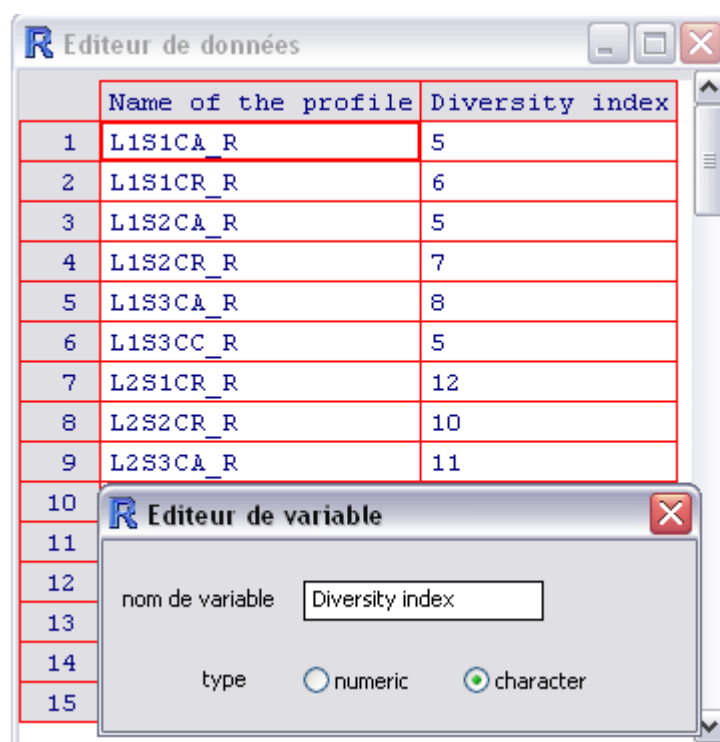
Fig 3: The main window of the program.

The status of each step of the fingerprint profile processing menu can also be checked (Fig 2).

Under the status part is located two consoles. The left one, named OUTPUT, gives the results of functions, whereas the right one, named WARNING, gives all warnings which could appear during the use of this program.

2- FILE MENU

2-1 Convert FSA files and import

Raw fingerprint profiles from an ABI Prism 310 or 3100 sequencer (Applied Biosystem) are available as FSA files which can be automatically converted to ASCII files and then loaded into STATFINGERPRINTS. The conversion step of this procedure needs the free program DataFilesConverter (Applied Biosystems) available at http://www.appliedbiosystems.com/support/software_community/tools_for_accessing _files.cfm. When running this procedure, the first step is to specify the folder where DataFilesConverter has been stored and then the folder where the FSA files are stored. During this step FSA files are converted into ASCII and stored in a folder named txt located in the FSA file folder. This step can take several minutes depending on the number of fingerprint profiles and your computer system. In the second step, channels containing the community profiles and the internal standards must be specified (Fig 4). This last step can easily be done using the HOW TO CHOOSE COMMUNITY AND INTERNAL STANDARD LOCATION button.
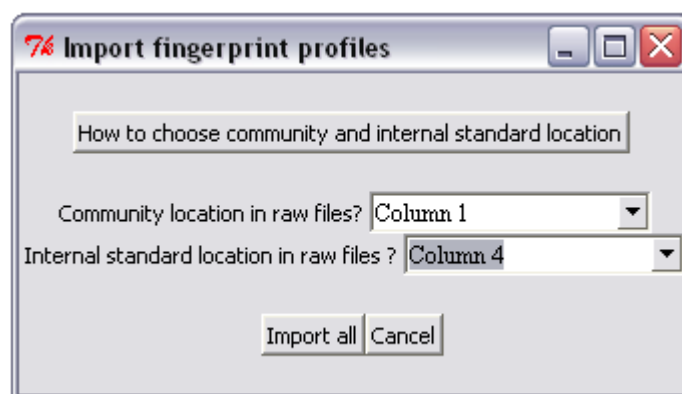
Fig 4: The window to convert and import FSA files.

2-2 Import fingerprint profiles as ASCII files

Fingerprint profiles in ASCII format can be imported with the STATFINGERPRINTS program as it can set the following parameters: field separator, decimal separator and occurrence of header. The easiest solution is to export files in CSV format using a text editor or spreadsheet program.

The first step is to select your ASCII files (at least two) and the second is to state the field separator, the decimal separator, the occurrence of headers and the columns containing the community and internal standard profiles (Fig 5). Specifying the columns containing the community and the internal standard profiles can easily be done using the HOW TO CHOOSE COMMUNITY AND INTERNAL STANDARD LOCATION button.
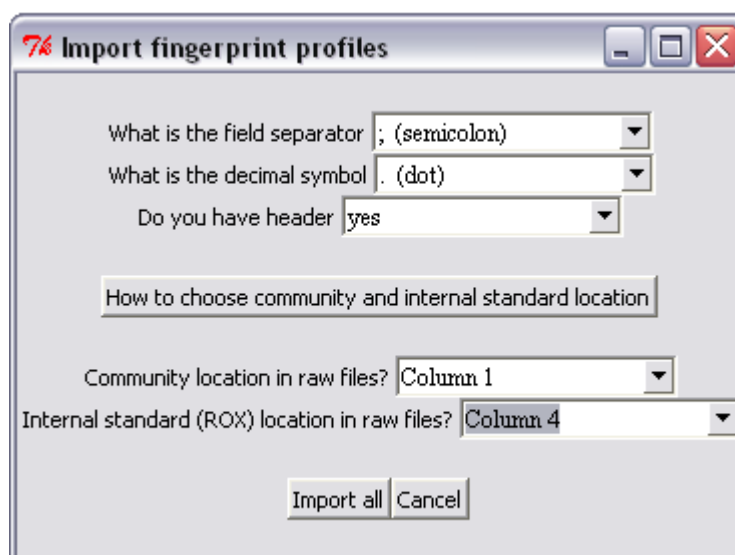

Fig 5: The window to import fingerprint profiles as ASCII files.

2-3 Import ecological table (ASCII)

Fingerprint profiles can also be imported using an ecological table as an ASCII file. An ecological table contains each microbial chromatogram in each row or column of

the table. Internal standards for each microbial community are not included in the table, so fingerprint profiles cannot be aligned.

To import an ecological table, first choose your file location and then complete the structure of the table (fingerprint profiles are in rows or columns), the field separator, the decimal separator and the occurrence of headers in columns and rows (Fig 6).
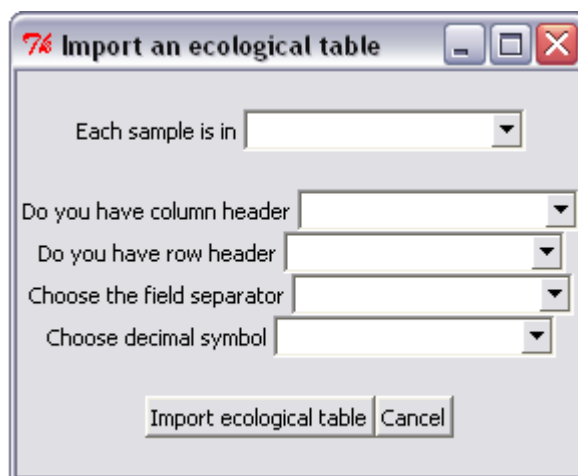

Fig 6: The window to import fingerprint profiles as an ecological table.

Information about the parameters of the ecological table can be found by opening the ASCII file with any text editor or spreadsheet program.

2-4 imported variables

Quantitative or qualitative variable tables in ASCII files are imported with this procedure. Be careful to separate your quantitative and qualitative data into two files as importing files containing both qualitative and quantitative variables is not supported. Special characters and special formats (bold, italic, underlined etc.) are also not supported. The missing value must be indicated with "NA" for not available. To import a table of variables, first select your file location and then specify the field separator, the decimal separator, the occurrence of headers and the type of the variables: qualitative (factor) or quantitative (parameter) (Fig 7).
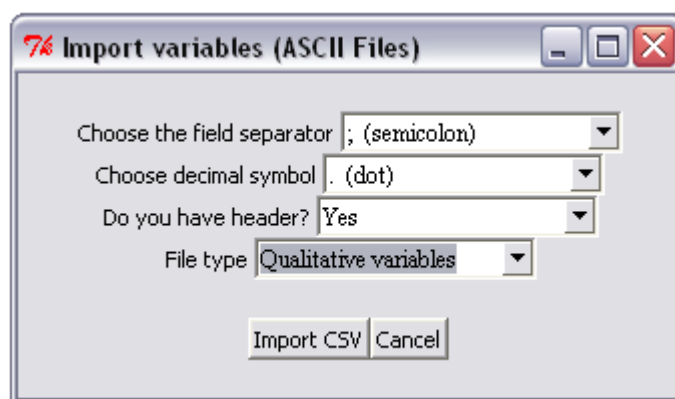

Fig 7: The window to import variables as ASCII files.

2-5 Load Project, save, save project as

We advise you to save your data regularly. All the data and objects created are stored in an Rdata file (see sections *8-1 Object invoked* and *8-2 Internal procedures*). "Save project as" and "save project" procedures are classical procedures to save a project; "Save project as" allows the file directory to be specified. The "Load" procedure allows a saved project to be loaded.

3- EDIT MENU

3-1 Change names of profiles

Names of fingerprint profiles can easily be changed using this procedure (Fig 8).
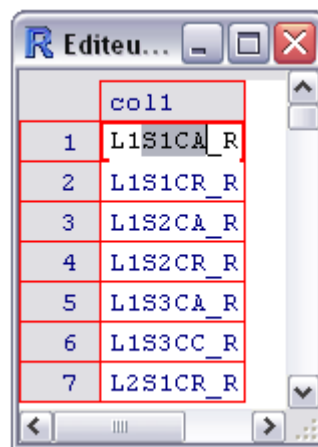

Fig 8: The window to change the name of fingerprint profiles.

First double clicks in the fingerprint profile; next write the new name. The name of the fingerprint profile will be immediately updated.

3-2 Add profiles to the project

This procedure allows two projects to be merged. Be aware that only projects saved with fingerprint profiles processed in exactly the same way are permissible (especially as regards the alignment step, see section *4-3 Align profiles one by one* for details).
This procedure requires three consecutive steps (Fig 9):

- 1: import the two projects. This means that you have to save your current project and create a new project with the fingerprint profiles to be added. The button LOAD TWO PROJECTS produces two consecutive exploratory windows to select your first and second project.
- 2: merge the two projects. Several warnings can appear to indicate if fingerprint profiles in the two projects have not been processed in the same way.
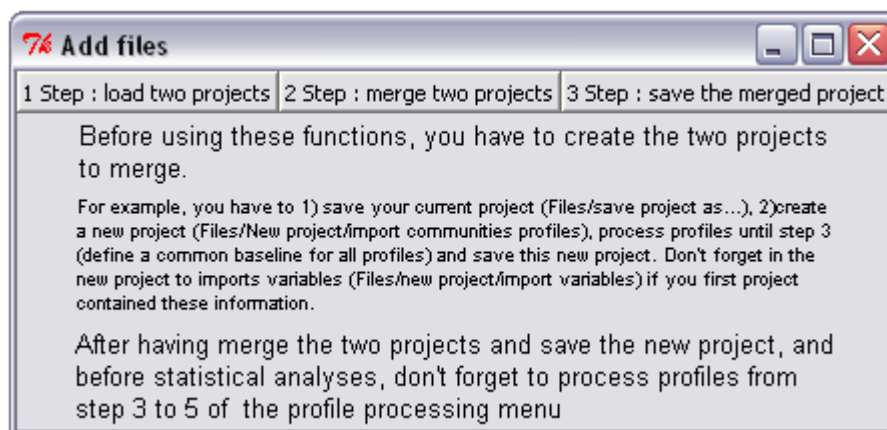- 3: save the new merged project.

Fig 9: The window to add fingerprint profiles to the project.

3-3 Delete profiles within the project

This procedure deletes one fingerprint profiles in your project (fig 10). Qualitative and quantitative variables corresponding to the deleted fingerprint profiles are also updated. This procedure should be use with care as the "delete fingerprint profile" will be completely erased from the project. We advise saving the current project in another file before using this procedure to make it possible to use the deleted fingerprint profiles if necessary.

Fig 10: The window to delete fingerprint profiles to the project.

3-4 Select profiles using levels of factor

This procedure allows a group of fingerprint profiles selected by the level of a factor to be deleted (Fig 11). Save the new project with a different name as the selected group of fingerprint profiles will be completely erased.
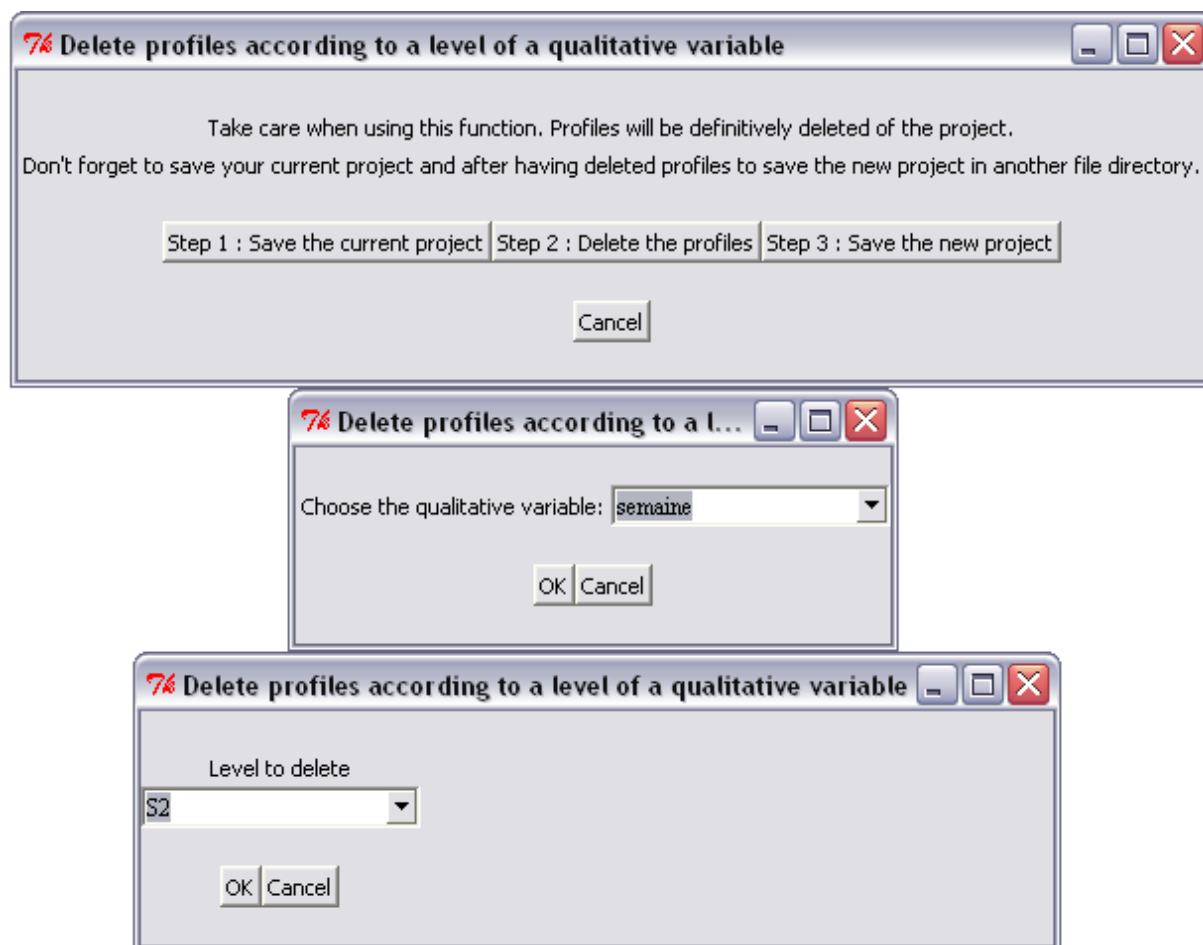

Fig 11: The window to select fingerprint profiles using level of factor.

4- PROFILE PROCESSING MENU

Before statistical analysis, fingerprint profiles have to be aligned together. Then several other treatments can be applied to the aligned fingerprint profiles to make them more comparable.

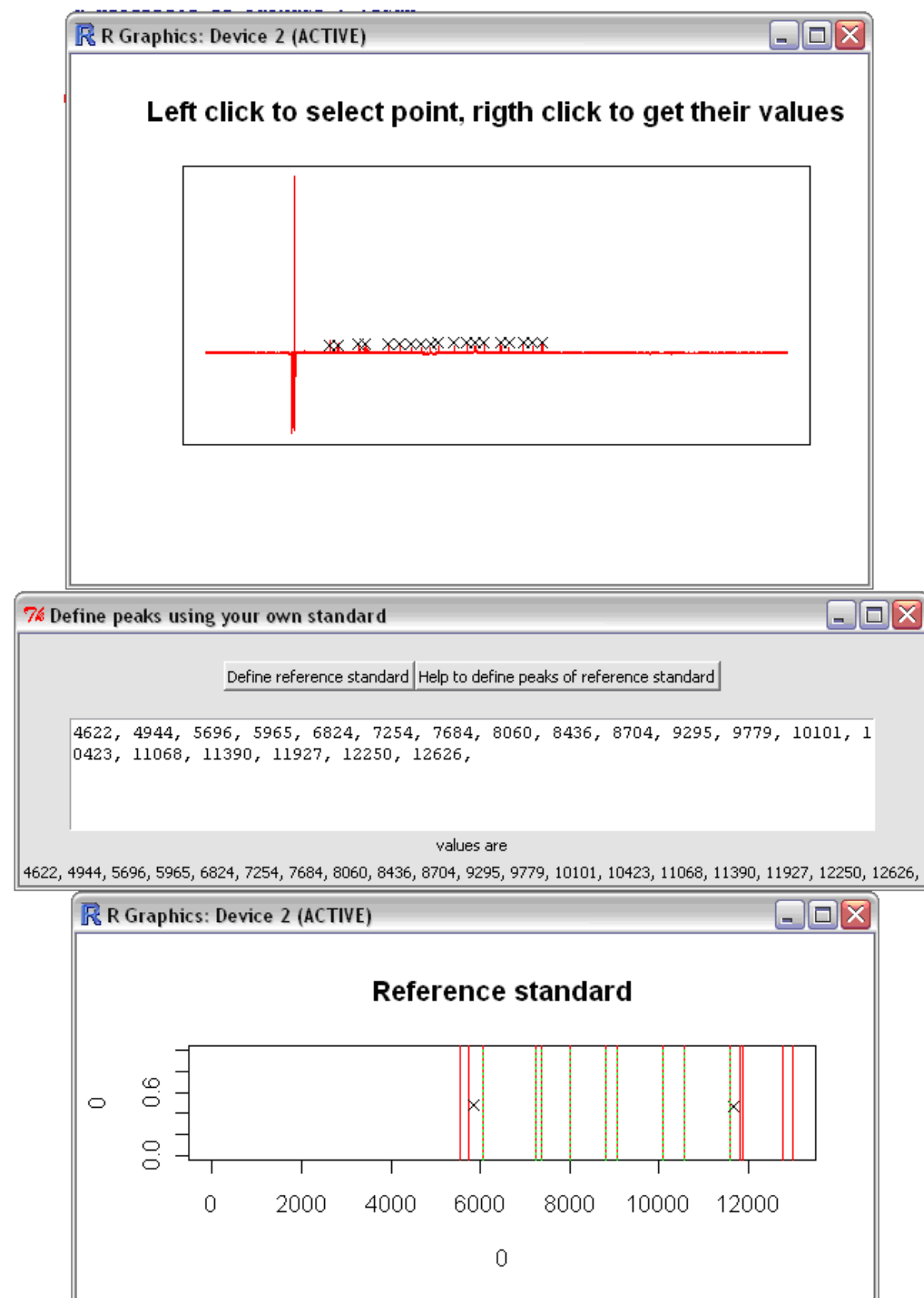### 4-1 Define peaks using your own reference standard

A reference standard is required to align fingerprint profiles as an internal standard of each fingerprint profile is aligned on this reference standard. This procedure consists of two steps.

- 1 (Fig 12 middle); the maximum abscissa of each peak of your reference standard must be entered in the dialog box. If necessary, maximum values of abscissa peaks can be found with the HELP TO DEFINE PEAKS OF THE REFERENCE STANDARD button. This procedure produces a signal corresponding to the

internal standard of your first file (Fig 12 top). Left clicks can then select peaks; once all peaks are selected, right click prints the abscissa values of the selected peaks in the dialog box (Fig 12 middle). Then validate the values by pressing the DEFINE REFERENCE STANDARD button.

- 2 (Fig 12 bottom); defines which peaks will be used to align the fingerprint profiles. Indeed, communities often spread out in a small area of the reference standard and thus alignment on the whole peaks of the standard is useless. To select the area where communities are located, left click to the left of the first peak and to the right of the last peak in this area. The peaks which will be used for alignment then appear as ticked red-green vertical lines.

Fig 12: The window to define a new reference standard. At the top is the procedure to get the abscissa values of peaks. These values then appear at the end of the tcltk window (middle). Then the user should specify peaks selected to perform alignment (bottom).

4-2 Use peaks of ROX HD400

Rox HD400 (Applied Biosystems) is an internal standard often added to samples when performing Capillary Electrophoresis Single-Strand Conformation Polymorphism (CE-SSCP). Abscissa values of peaks of the ROX HD400 are automatically loaded with this procedure but peaks of the ROX HD400 used for the alignment must be defined as described in the second step of the section 4-1 (*Define peaks using your own reference standard*).
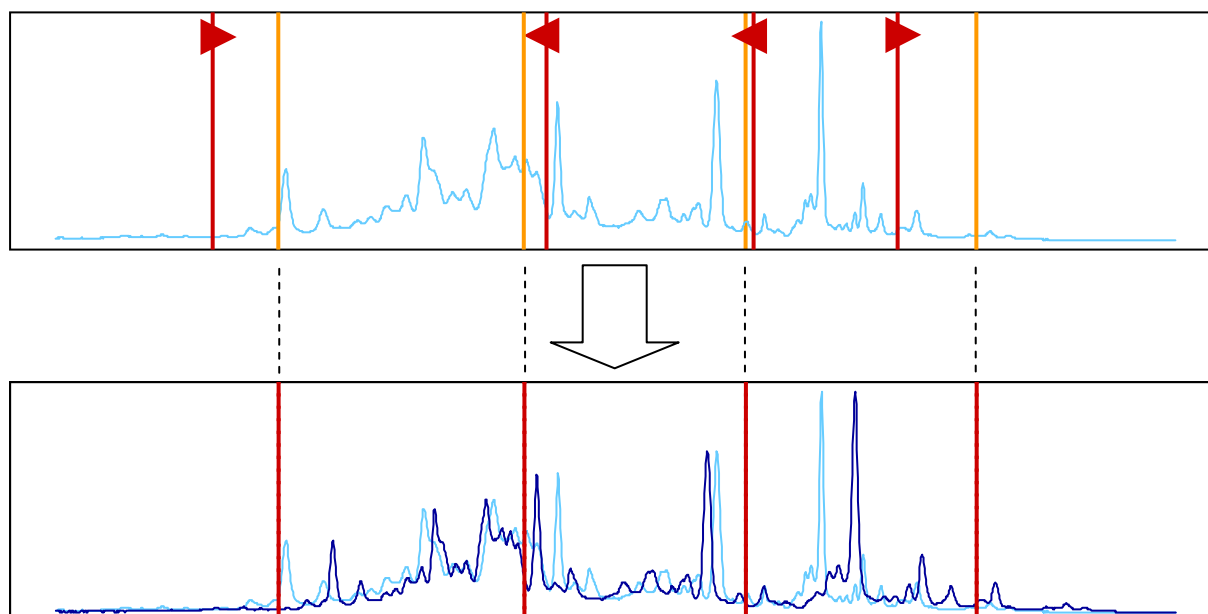
4-3 Align profiles one by one

Fig 13: Scheme of the alignment. Peaks of the reference standard are in orange. The microbial community fingerprint profile and its internal standard are in blue and red respectively. Alignment consists of aligning the internal standard with the reference standard and applying the same transformations to the microbial community fingerprint profile.

Peaks of the reference standard must be defined before using this procedure (see sections *4-1 Define peaks using your own reference standard* or *4-2 Use peaks of ROX HD400*).

This procedure uses an algorithm of cubic spline interpolation of fingerprint profiles which consists of fitting an exact cubic spline through the four scans at each end of the scans (Forsythe et al, 1977).

Basically, the algorithm aligns peaks of the internal standard of the processed fingerprint profile with those of the reference standard and then it applies the same computed transformations to the community signal of the processed fingerprint profile (see Fig 13). Therefore, alignment is only a matter of computing each fingerprint profile and must be applied in the same way for the entire set of fingerprint profiles.

Alignment consists of 5 consecutive steps:

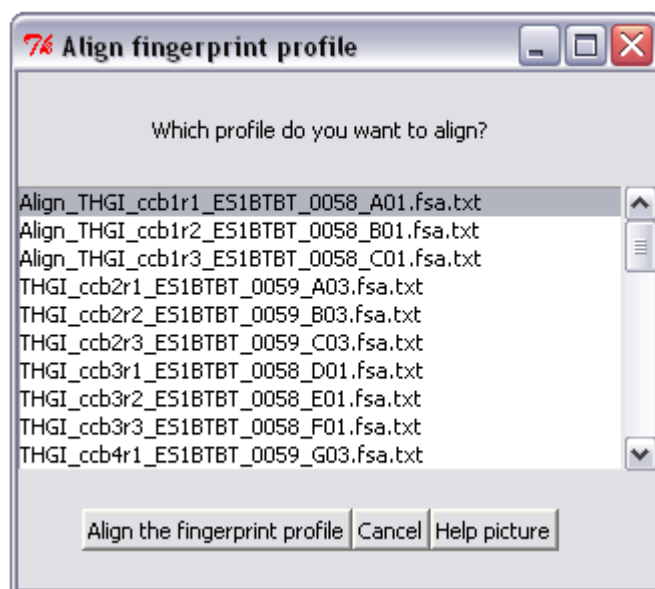- 1: select the fingerprint profile to be aligned (Fig 14).

Fig 14: The window to align a fingerprint profile.

- 2: in the open R graph, mark the area corresponding to the peaks previously defined in the reference standard (see sections *4-1 Define peaks using own reference standard* or *4-2 Use peaks of ROX HD400*). To do this, use two left clicks as before on either side of the area to be marked.
- 3: in the following R graph, select the area containing the same peaks as those defined in the reference internal standard by clicking before the first peak and after the last peak corresponding to those defined in the reference standard (Fig 15). The y-value of the first click must be less than the maximum y-values of all peaks.
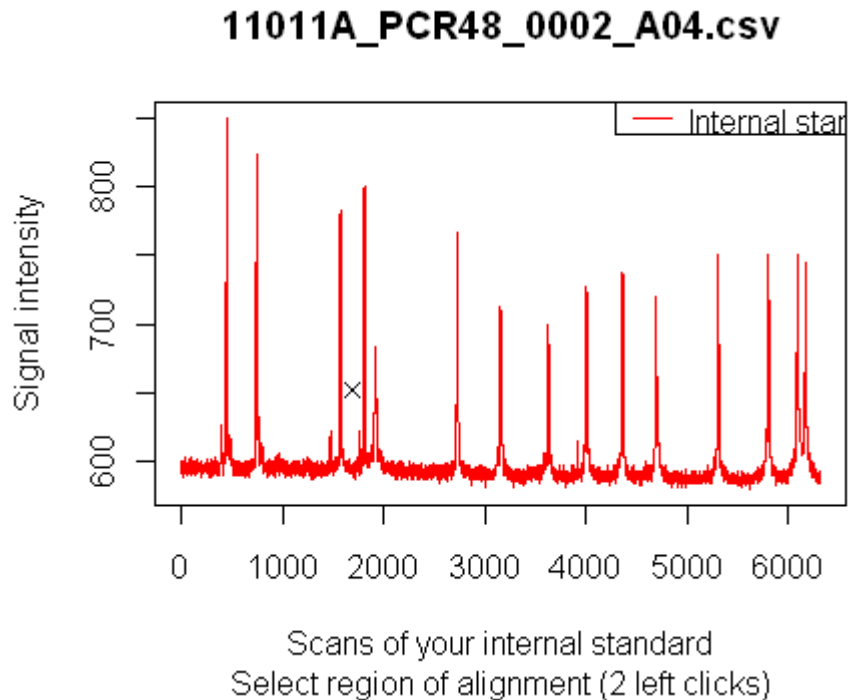
11011A_PCR48_0002_A04.csv

Fig 15: The selection of the area of the internal standard of the fingerprint profile containing the same peaks as those defined in the reference standard.

- 4: false peaks due to artefacts often occur along the internal standard. This step allows these false peaks to be deleted by left clicking before and after each false peak. Once all false peaks have been selected, go to the next step using a right click. If no false peak is present, go to the next step by making two left clicks in an area containing no peaks, and validate by a right click (Fig 16).
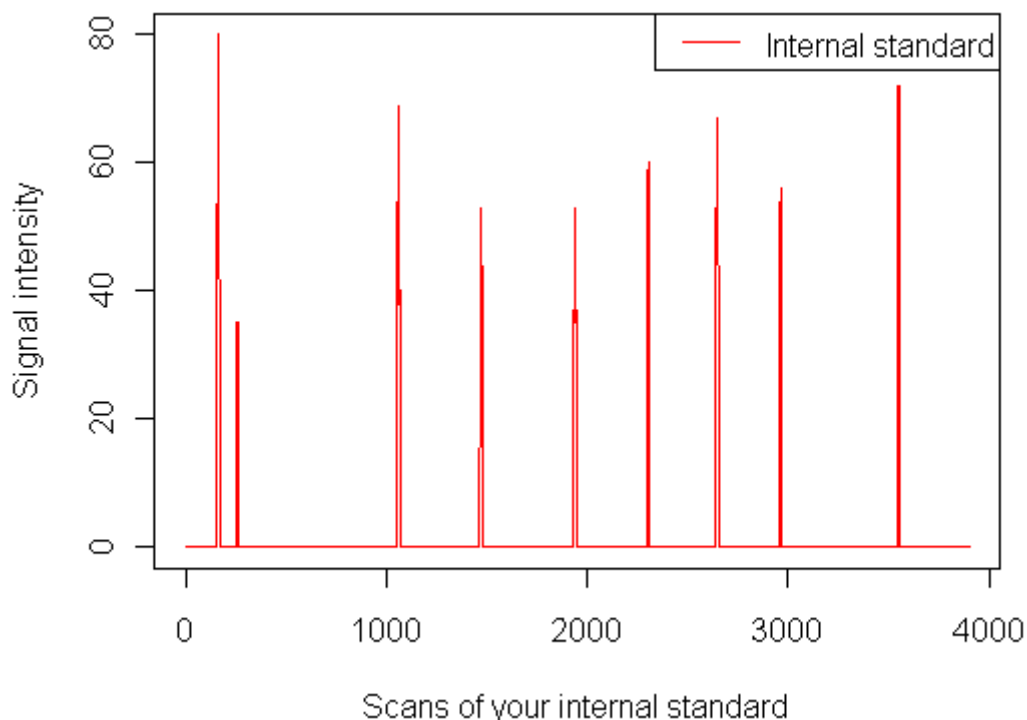

Fig 16: Suppression of a false peak of the internal standard of a fingerprint profile (artefact).

- 5: the open R graphic window represents the result of the alignment. The number of detected peaks on your fingerprint profile must correspond to the number of peaks of the reference standard in order to be aligned. If these two numbers of peaks differ the alignment is not performed and a warning box appears. A wrong number of peaks in the fingerprint profile is often due to either a different selected location from that used for the reference standard (see step 3 of this procedure and sections 4-1 *Define peaks using your own reference standard* or *4-2 Use peaks of ROX HD400*) or to the occurrence of undeleted false peaks (see fourth step of this procedure). Once correctly aligned, "Align_" will be added to the name of the fingerprint profile in the dialog box of the procedure (Fig 14).
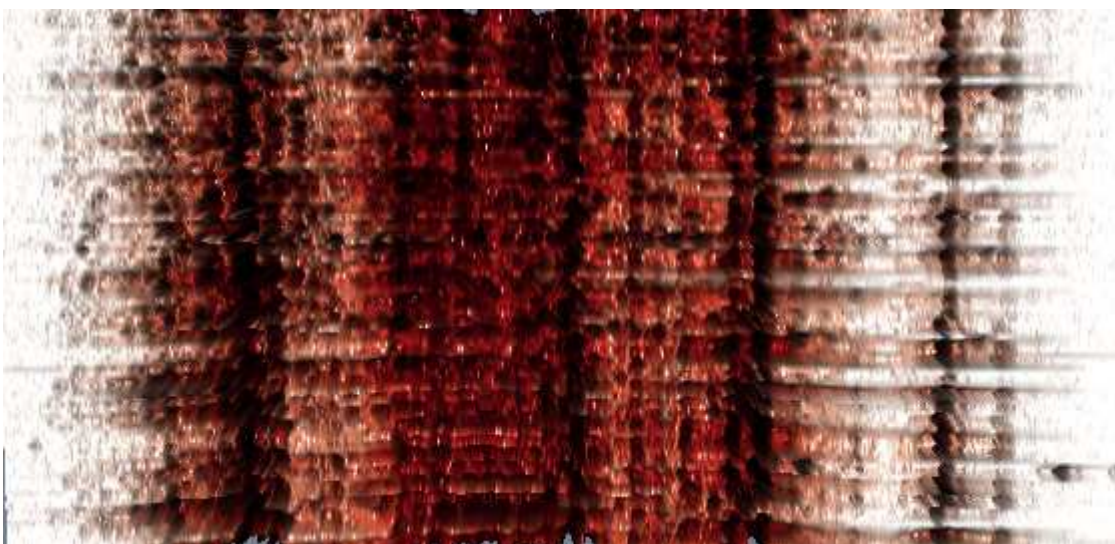
4-4 Check quality of the alignment

Fig 17: The rotation of this 3D plot of the whole fingerprint profiles allows the quality of the alignment to be checked.

This procedure plots all the aligned fingerprint profiles in three dimensions (Fig 17). The graph can be rotated using successive left clicks and thus alignment can be visually checked.

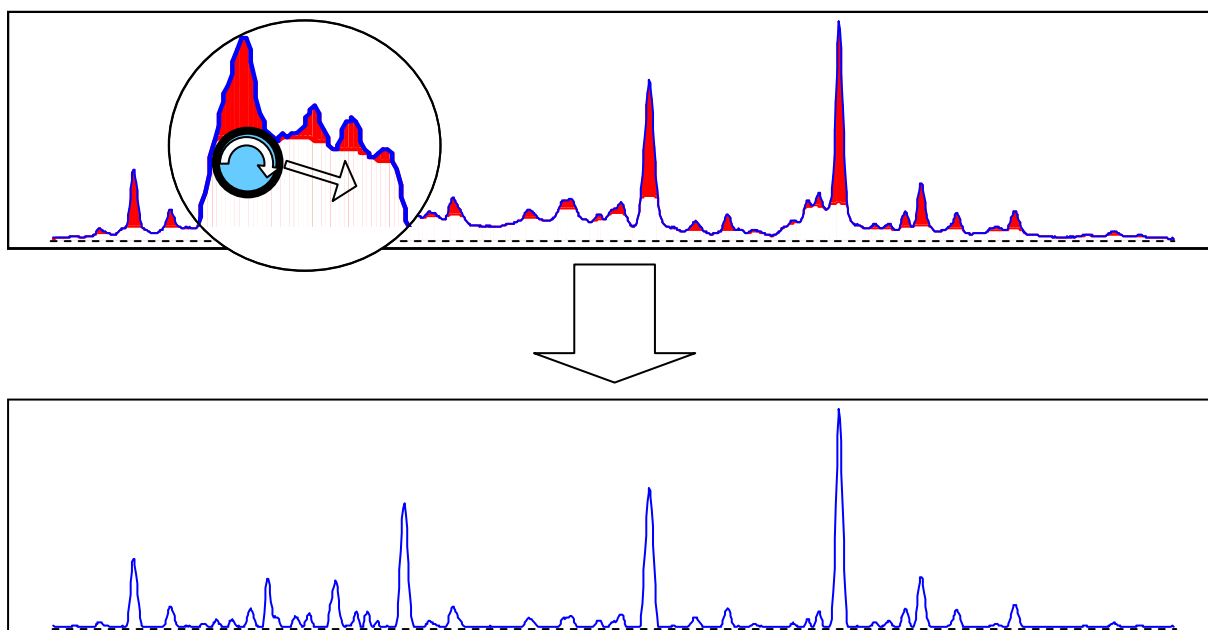## 4-5 Delete background under the profiles

Fig 18: Scheme of the algorithm of the rollball to delete the background under peaks of fingerprint profiles.

Optionally, backgrounds under all fingerprint profiles can be eliminated using the "rollball" algorithm. This removes the area under the trajectory of a virtual ball, rolling under the signal of the community of the fingerprint profile (Fig 18). A threshold must be specified to determine the radius of the virtual ball to delete more or less background (Fig 19).
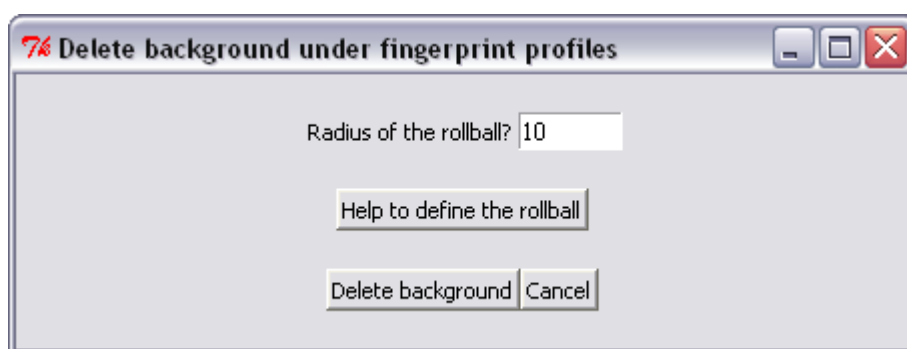

Fig 19: The window to delete the background under peaks of fingerprint profiles.

The most suitable radius for your fingerprint profiles can easily be defined using the HELP TO DEFINE THE ROLLBALL button. This procedure plots your first fingerprint profile before and after deleting the background using the specified radius of the rollball.

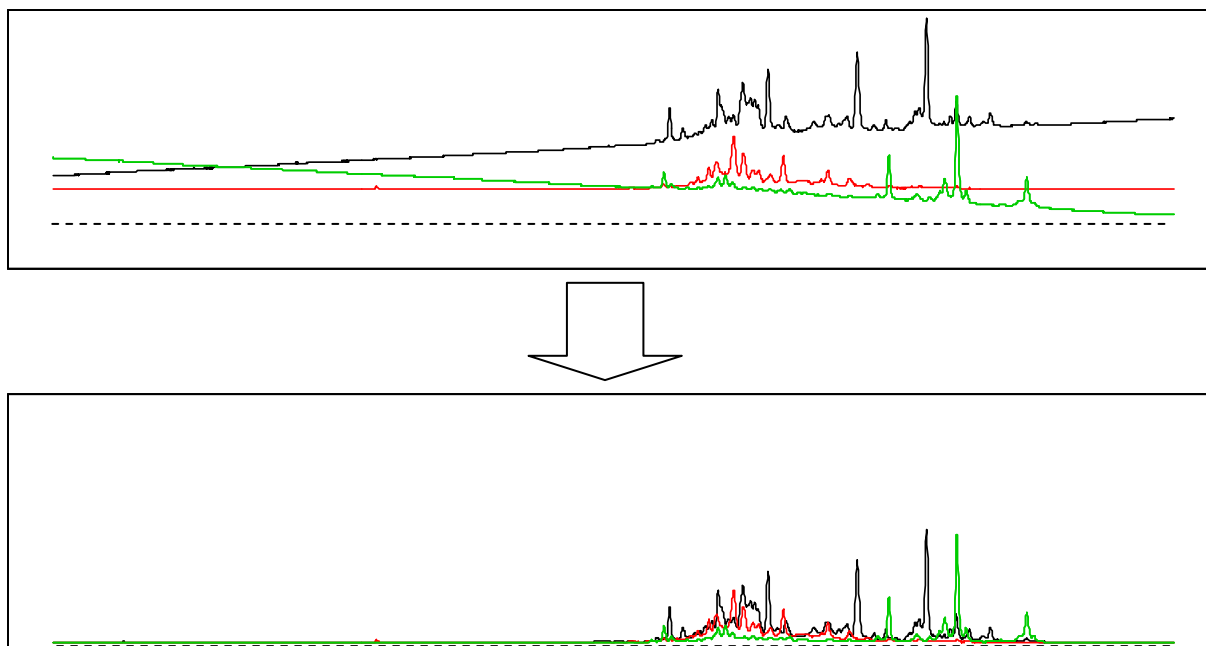## 4-6 Define a common baseline for all profiles

Fig 20: Scheme of the procedure to define and to direct the baseline in an equal horizontal way.

As illustrated in Fig 20, fingerprint profiles often have their baselines non-aligned (not the same y-values) or some fingerprint profiles (red or green signal in Fig 20) are not perfectly parallel to the abscissa axis. To align as well as establish a horizontal baseline for all fingerprint profiles, two left clicks are needed, the first just before the beginning of the signal of the microbial community and the second just after it.

4-7 Define the range of the profiles

This procedure allows only the area corresponding to the microbial community in the fingerprint profiles to be kept. On the R graphic window, left click before and after the area of interest of all profiles (Fig 21).
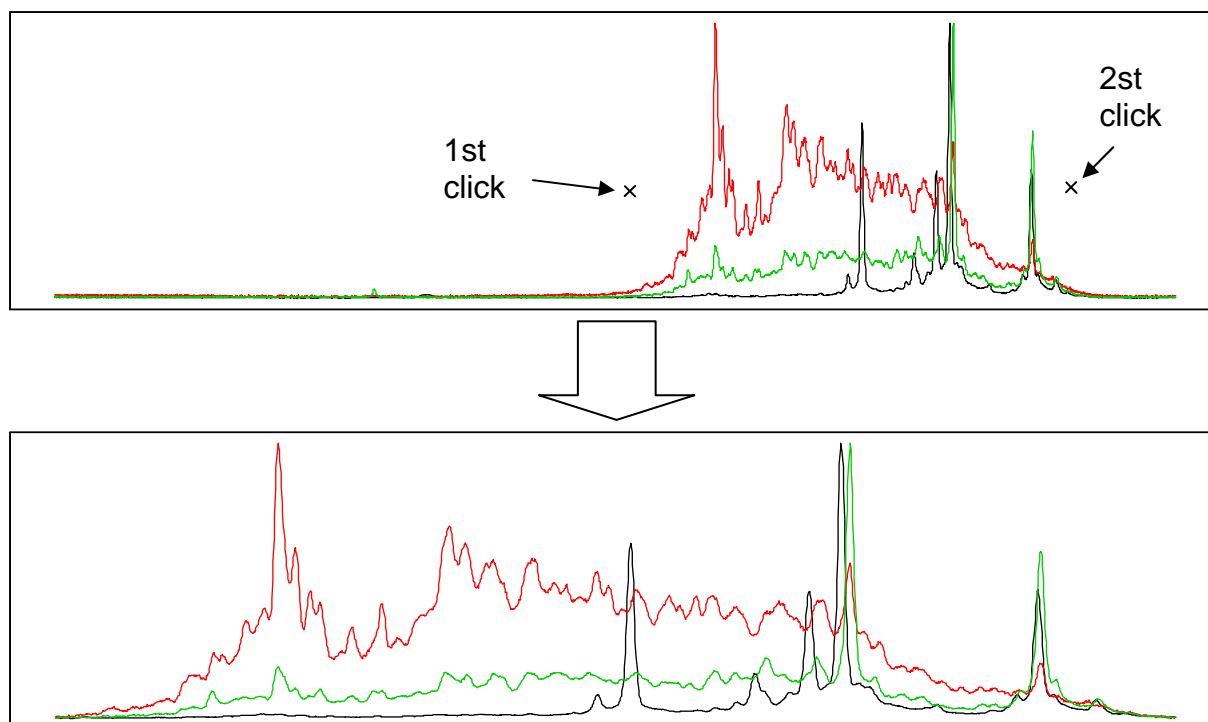
Fig 21: Scheme of the procedure to define the range around the community of the fingerprint profiles.

4-8 Rebuild peaks of defective profiles (optional)

Care must be taken when using this procedure and it is preferable to restart the laboratory work on the fingerprint if possible rather than using it.
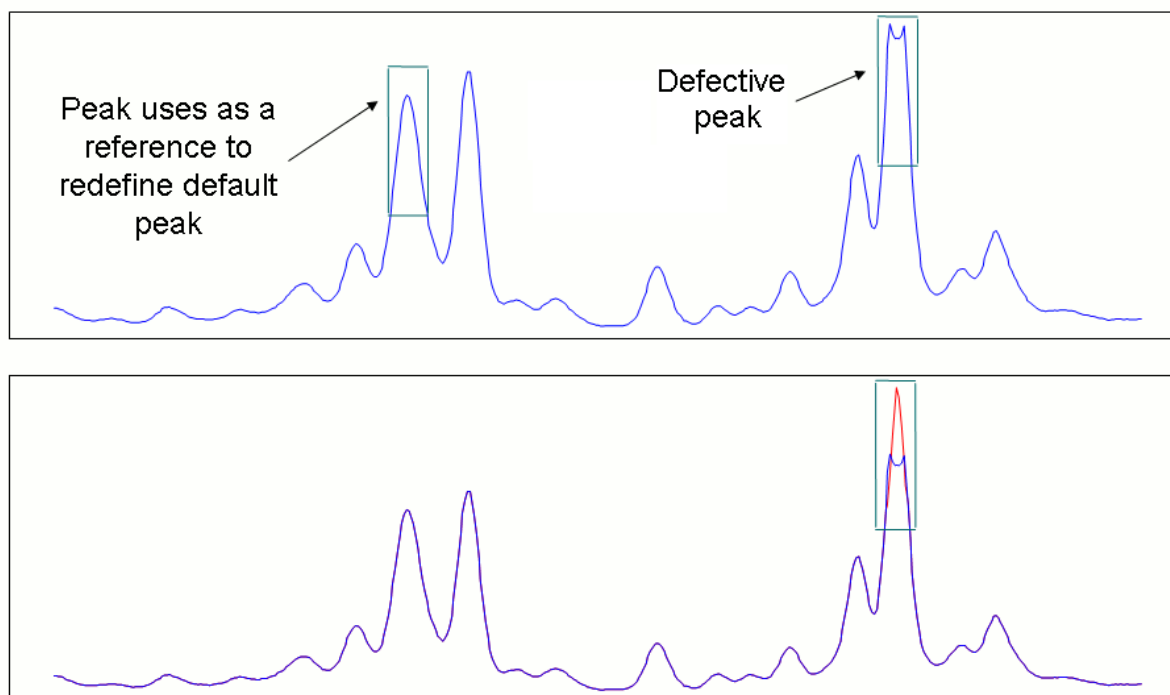

Fig 22: Scheme of the result of the procedure to rebuild defective peaks.

This procedure consists of six steps:

- 1; select the fingerprint profile with one or more defective peaks and press the COMPUTE PEAK MODIFICATION button (Fig 23). Defects on peaks of fingerprint profiles can be seen using the procedure "plot fingerprint profiles in 2 dimensions" (see section *5-1 Plot profiles in 2 dimensions*).
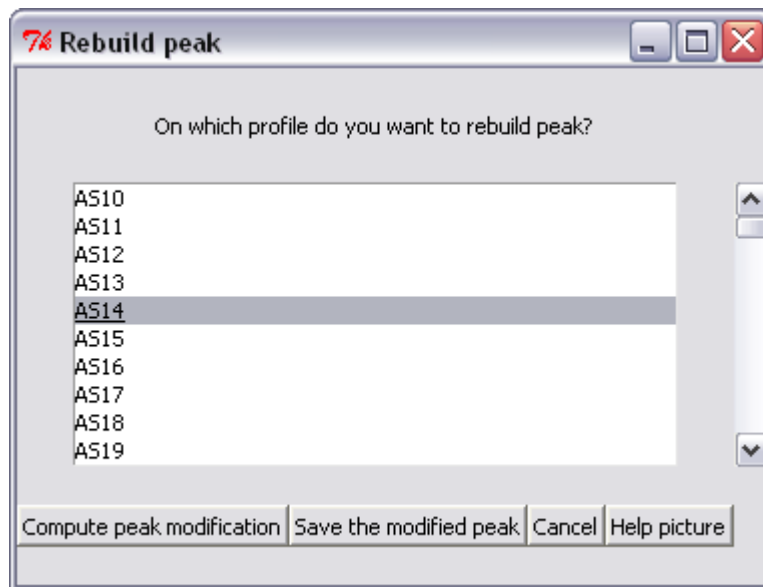

Fig 23: The window to modify defective peaks.

- 2; select (using two left clicks) an area presenting a peak similar to the defective one. This peak will represent the reference peak.
- 3; select precisely the start and end of the reference peak using two left clicks (Fig 22). The curve equation of this reference peak will be defined as a model to rebuild the defective peak.

Validate the area of the reference peak (rigth click)

Fig 24: Plot resulting from the automatic selection of a peak

- 4; select the area of the defective peak using two left clicks.
- 5; select precisely the start and end of the defective peak using two left clicks.
- 6; inspect the quality of the new peak compared with the old one. If you are satisfied, save it by pressing the SAVE THE MODIFIED PEAK button (Fig 21): otherwise restart the procedure.

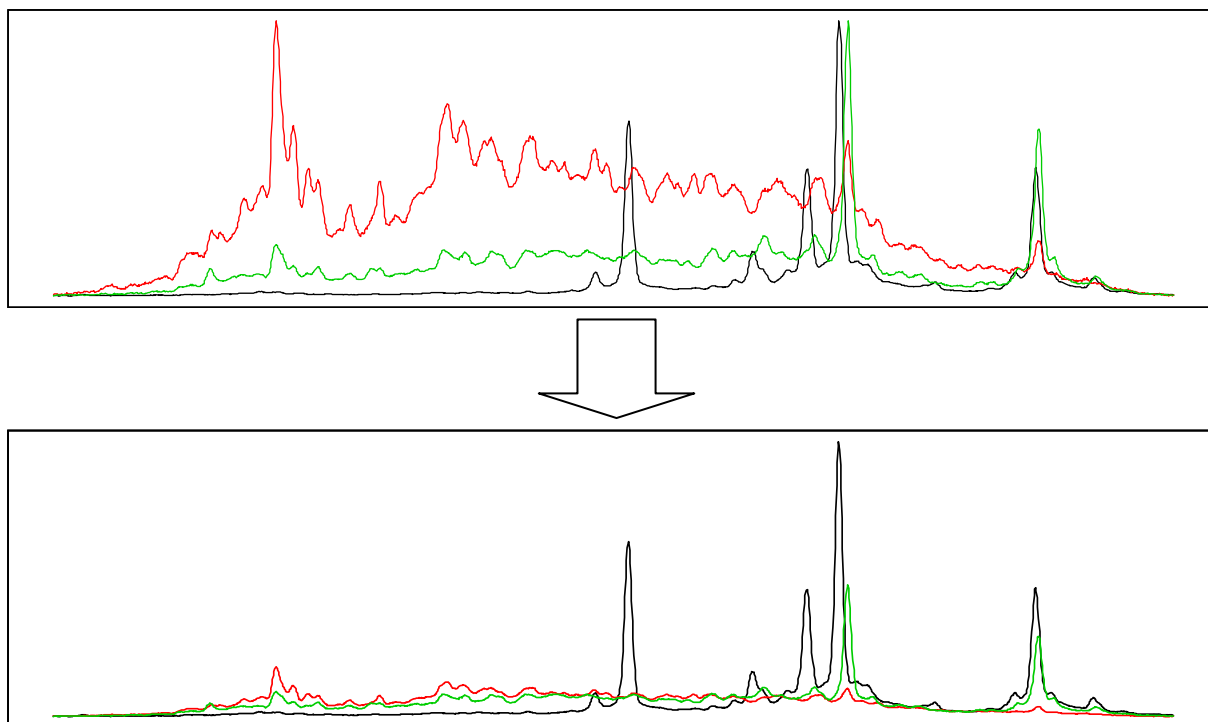## 4-9 Normalise the area under the profiles

Fig 25: Scheme resulting from the normalisation of the fingerprint profiles.

As illustrated by Fig 25, the area under each fingerprint profile is usually not the same. To efficiently compare fingerprint profiles, it is strongly advised to normalise them. Three different algorithms are provided for normalisation (Fig 26):

- Normalise area under curve so that the new area is equal to one, ignoring negative values (Fig 26 top).
- Convert all negative values into 0 and then proceed to alignment so that the new area under the curve is equal to one (Fig 26 middle).
- Take the minimum y-value of the curve and subtract the absolute of this value from all values within the curve and then proceed to alignment so that the new area under curve is equal to one (Fig 26 bottom).
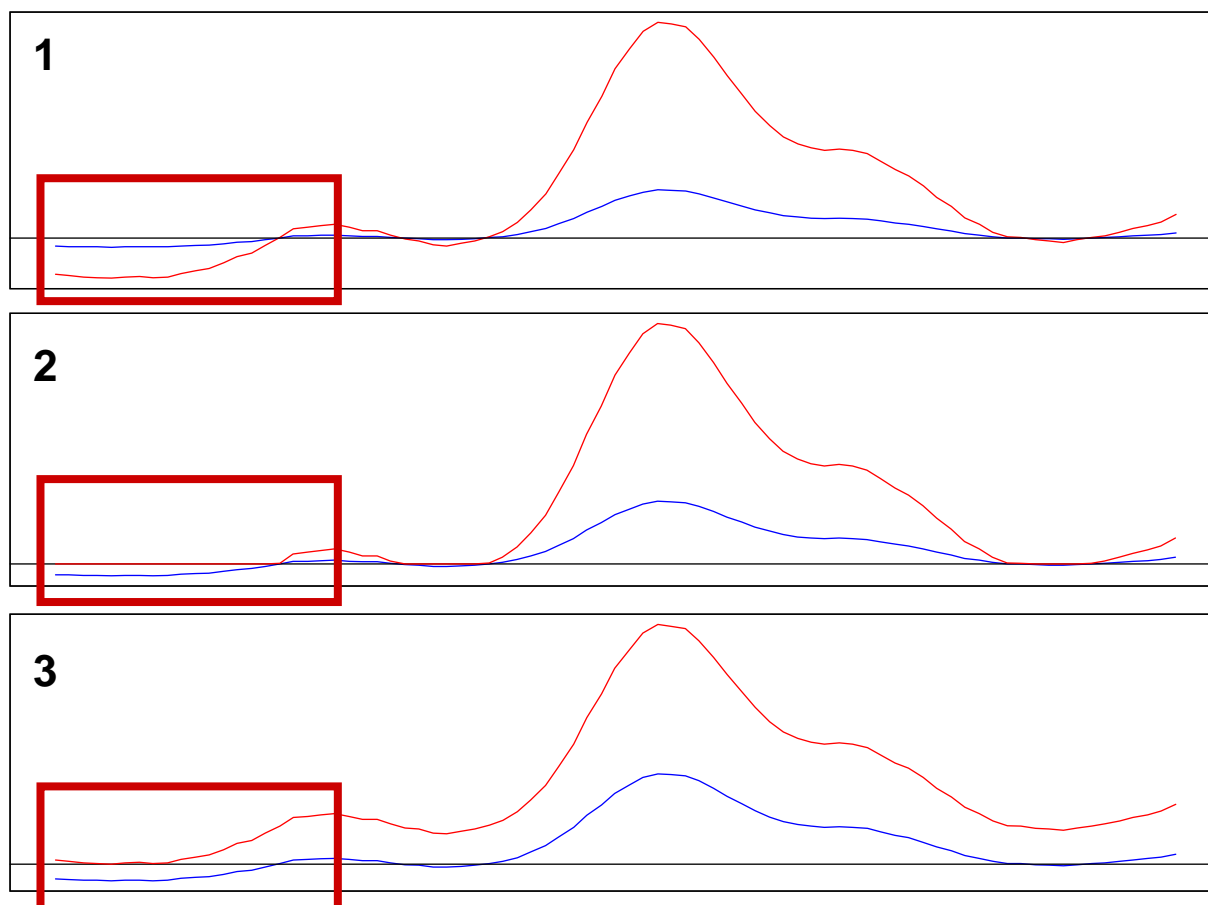
Fig 26: Scheme of the three different algorithms available to normalise fingerprint profiles.

4-10 Transform profiles into presence/absence profiles

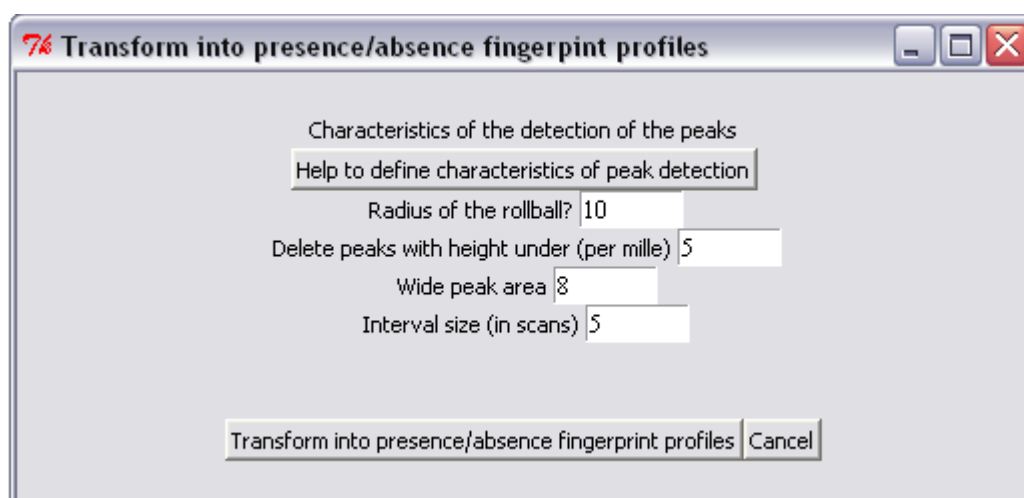This procedure transforms quantitative fingerprint profiles into binary fingerprint profiles.

Fig 27: The window to transform the fingerprint profiles into presence/absence fingerprint profiles.

Binary fingerprint profiles are required when using binary proximity measures (Jaccard, Dyce-Sorensen, Ochiai, Steinhaus). It means that each scan value of the fingerprint profiles is transformed into 1 or 0 according to whether it is located within a peak or not. The algorithm used to detect peaks can be fully parameterized to best fit each fingerprint profile set (Fig 27). Parameterized features are: the radius of the rollball (see section *4-5 Option: delete background under fingerprint profiles*), the threshold, the wide peak area, and the interval size (for details of the 3 last features; see section *6-1 Compute diversity index*). The values of these features can easily be defined using the HELP TO DEFINE CHARACTERISTICS OF PEAK DETECTION button (see section *6-1 Compute diversity index* for more details of this help procedure).

Fig 28: Plot showing the transformation of fingerprint profiles into presence/absence fingerprint profiles. Each fingerprint profile is in a row with black horizontal segments representing areas where peaks occurred.

As a result, the R graphic window displays a plot with each fingerprint profile drawn as a dotted line, the black horizontal segments representing detected peaks (value equal to 1; Fig 28).

Once this procedure has been executed, the previous fingerprint profiles with the quantitative information can be easily recovered by executing the normalisation procedure (see section *4-9 Normalise area under the profiles*)

5- PLOT MENU

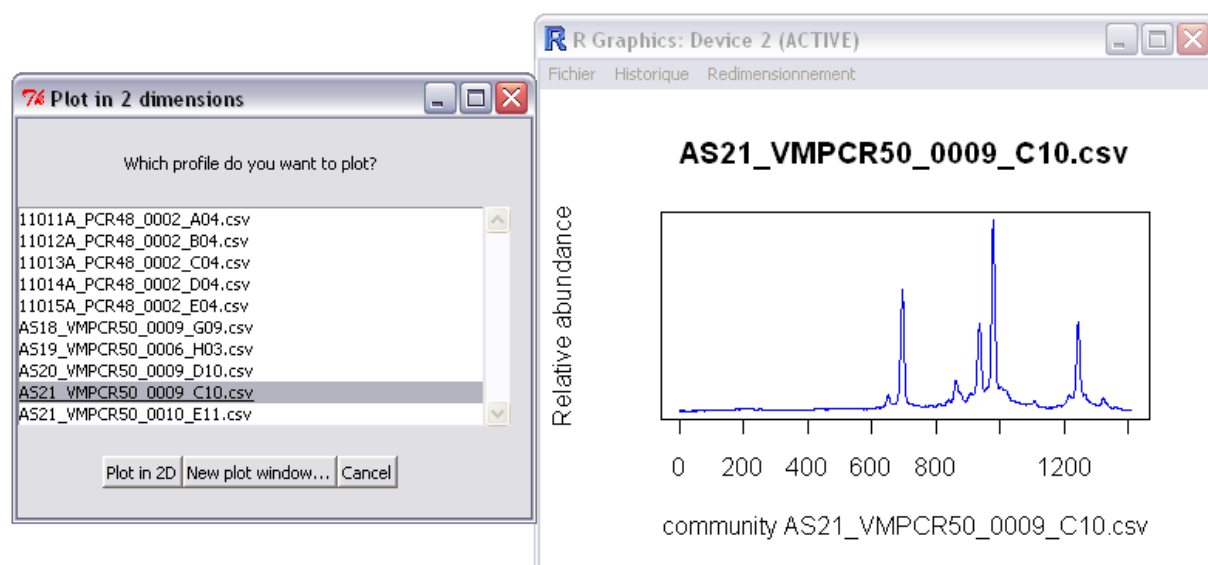### 5-1 Plot profiles in 2 dimensions

Fig 29: Example of a 2-dimensional fingerprint profile plot.

This procedure plots the community signal of a single fingerprint profile (Fig 29). It automatically detects the last step of fingerprint profile processing (from the importation to the transformation into presence/absence fingerprint profiles) and thus always plots the selected fingerprint profile in its latest state.

5-2 Plot profiles in 3 dimensions

Fig 30: Example of the whole fingerprint profile plotted in 3 dimensions.

This procedure plots the community signal of all the fingerprint profiles in 3 dimensions (Fig 30). It automatically detects the most recent step of fingerprint profile processing (from the importation to the transformation into presence/absence fingerprint profiles) and thus always plots fingerprint profiles in their most recent state. For the alignment, the procedure plots the aligned fingerprint profiles only if all

fingerprint profiles are aligned. The plot can be rotated using left clicks and zoomed into using right clicks. The plot can be saved as PNG using the SAVE PICTURE button.

5-3 Plot saved nMDS or PCA in 2 dimensions

This procedure can be used only after having performed and saved an ordination (see sections *7-1 Non-metric Multidimensional Scaling* and *7-2 Principal Components Analysis*).
This procedure allows results of ordination already computed (nMDS or PCA) to be plotted in two dimensions. The following features can be specified: selection of axes, labels, colour of points according to a qualitative variable, regression line according to a quantitative variable (Fig 31).
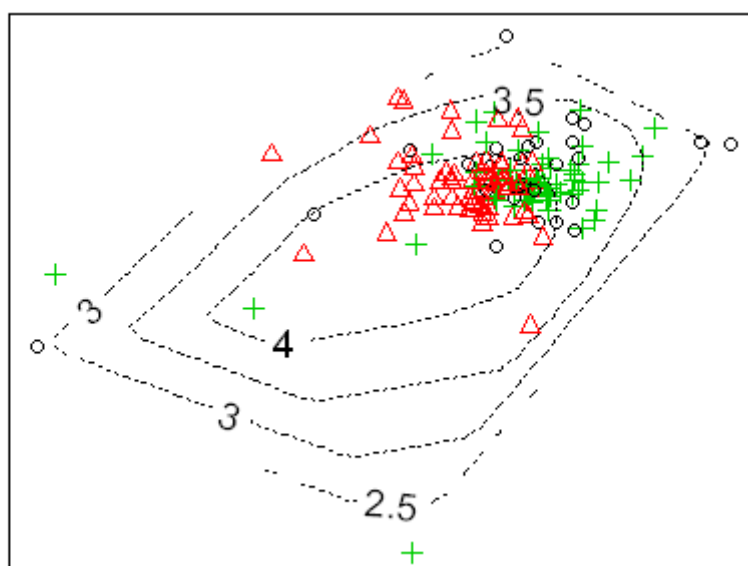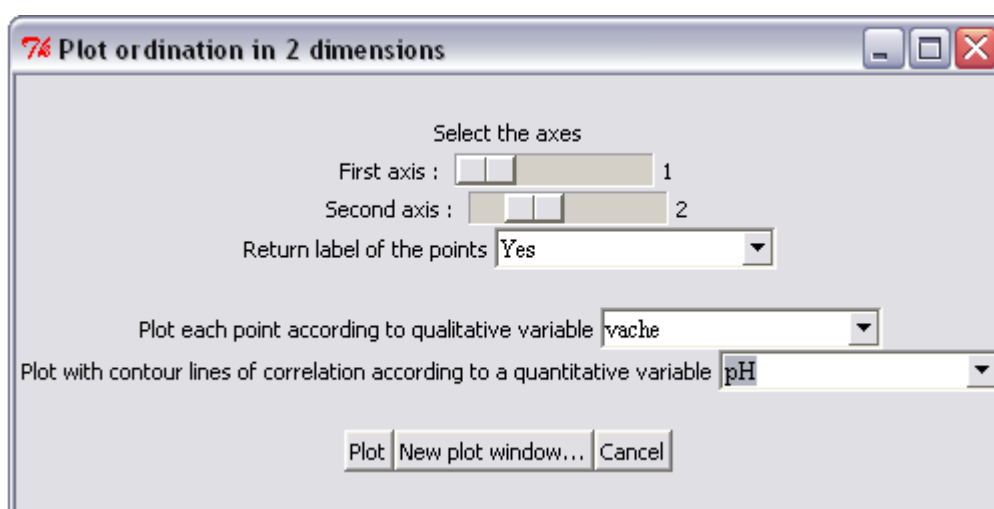

Fig 31: The window to plot ordination with advanced tools in 2 dimensions (top) and the resulting plot with contour regression lines and points coloured according to a qualitative variable (bottom).

The first step is to select a saved ordination and the second is to specify features of the plot.

5-4 Plot saved nMDS or PCA in 3 dimensions

This procedure can be used only after having performed and saved an ordination (see sections *7-1 Non-metric Multidimensional Scaling* and *7-2 Principal Components Analysis*). It can plot a computed ordination (nMDS or PCA) in three dimensions with dynamic control. It includes the following features: selection of axes, labels, colour of points according to a qualitative variable (Fig 32).
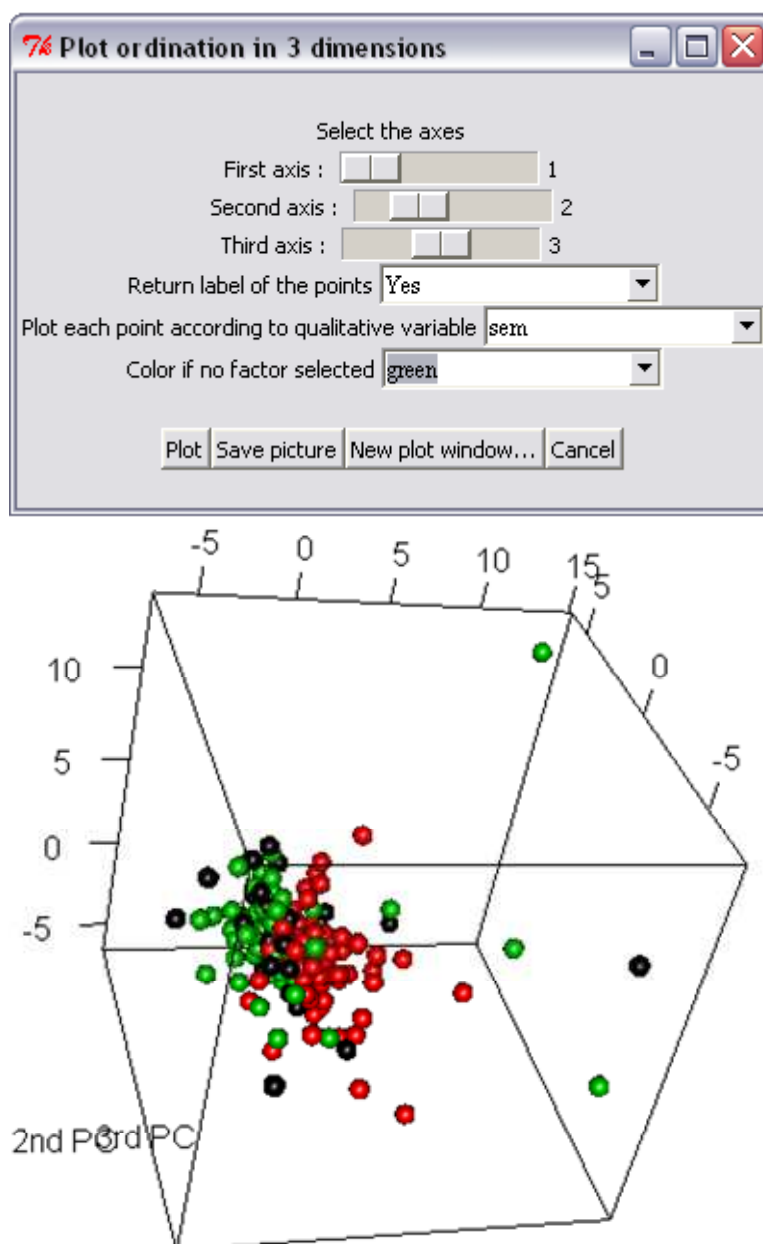

Fig 32: The window to plot ordination with advanced tools in 3 dimensions (top) and the resulting plot with points coloured according to a qualitative variable (bottom).

First select a saved ordination and then specify features of the plot. The plot can be saved as PNG using the SAVE PICTURE button.

6- UNIVARIATE STATISTICS: DIVERSITY INDEX MENU

6-1 Compute diversity index

Estimating a diversity index consists of summarizing a complex community represented by a fingerprint profile as a single value.

Fig 33: The window to compute diversity indexes.

Various diversity indexes can be calculated by taking into account either the number of peaks of the fingerprint profile or the number of peaks and their relative abundances (area or height under each peak of the fingerprint profile; Fig 33). The following diversity indexes are available (Magurran, 2004):

- Peak Number S (often named Richness)

- The minus logarithm of Simpson $D = -\log \sum \left(\dfrac{a_i}{\sum a_i} \right)^2$ where a_i is the relative abundance of each peak. If normalization is performed, the minus logarithm of Simpson is calculated as $D = -\log \sum a_i{}^2$. This ranges from 0 (a single peak) to infinity (an infinite number of peaks of equal abundance).

- One minus Simpson $D = 1 - \sum \left(\dfrac{a_i}{\sum a_i} \right)^2$

   It ranges from 0 (a single peak) to 1 (an infinite number of peaks of equal abundance).

- The Shannon index (entropy) $H = -\sum a_i \times \log a_i$ where a_i is the relative abundance of each peak. It varies from 0 for communities with a single peak to high values for communities with many peaks, each with little abundance.

- Buzas and Gibson's evenness $\dfrac{\exp^{-\sum a_i \times \log a_i}}{S}$ where a_i is the relative abundance of each peak and S is the number of peak.

- Equitability $\dfrac{-\sum a_i \times \log a_i}{\log S}$

Before computing diversity, the procedure needs to detect peaks and their sizes. The algorithm for peak detection can easily be parameterized to best fit to your fingerprint profiles.

The following parameters can be specified (Fig 33):

- the radius of the rollball (see section *4-5 Option: delete background under fingerprint profiles*).
- a threshold below which peaks are deleted.
- the width of the detected peaks.
- the interval size. This feature fixes the scanning range within which the maximum y value is sought. The smaller the value, the better is the result but, it is more time-consuming.
- the method of calculation. There are 2 methods for calculating abundance, one using the maximum height of peaks and the other using area under peaks.

The HELP TO DEFINE CHARACTERISTICS OF PEAK DETECTION button helps to choose the values of parameters (Fig 34). It quickly plots the result of a single fingerprint profile with the given values of the parameters and thus allows these values to be fitted.

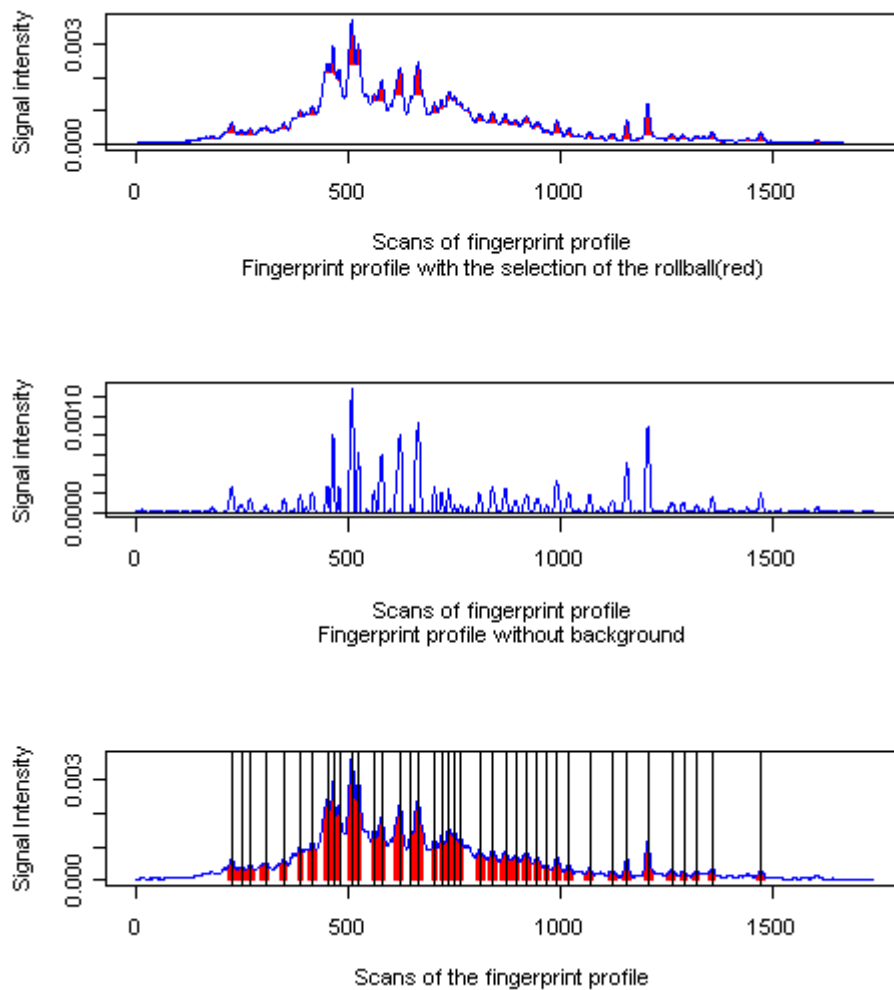**Fig 34: Plot showing the different steps of a diversity index calculation for each fingerprint profile.**

6-2 Descriptive statistics

This procedure computes basic descriptive statistics (mean and standard deviation) of either a diversity index or a quantitative variable according to the levels of a qualitative variable (Fig 35).
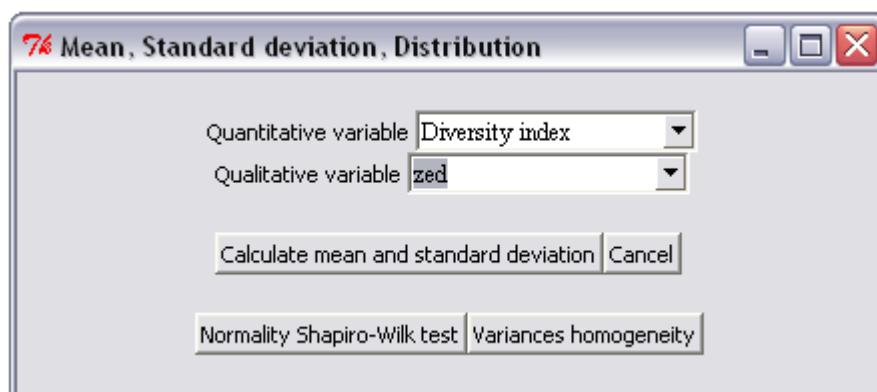

Fig 35: The window to compute basic descriptive statistics.

The normality of the distribution of a quantitative variable and the homogeneity of variance (necessary assumptions for ANOVA) are also provided.
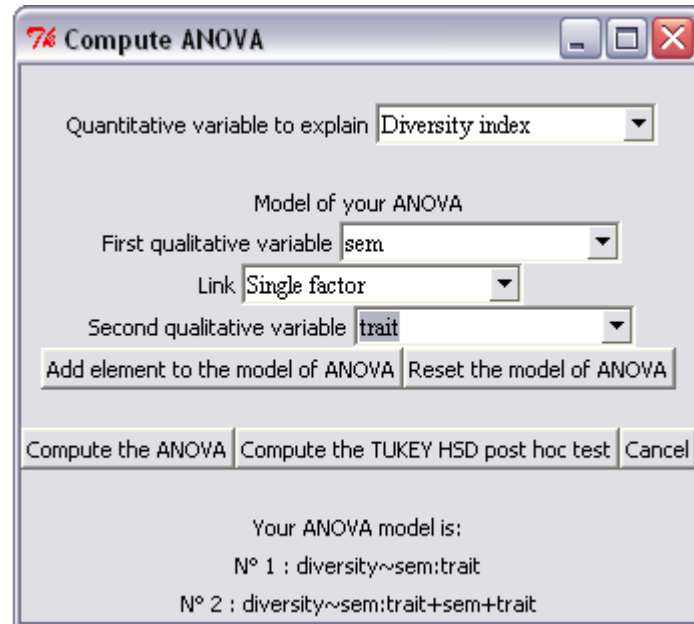
6-3 Multifactor ANOVA

Fig 36: The window to define a design and to compute ANOVA and Tukey HSD test.

Multifactor ANOVA is a statistical procedure for testing the null hypothesis that a quantitative variable has the same mean across each of several factors, and that there are no dependencies (interactions) between these factors. This procedure computes a type 2 ANOVA. The samples are assumed to be roughly normally distributed and to have similar variances. If the sample sizes are equal, these two assumptions are not critical.

The ANOVA model must be built by adding two factors and their link (independently, interaction, independently + interaction) to the model and by doing it again until the complete model is achieved. In the example of Fig 36, the first model presents the interaction between "sem" and "trait" factors, the second one presents the same interaction plus the two factors independently. If the model has an error, it can be reset using the RESET THE MODEL OF ANOVA button. Once the model is complete, use the COMPUTE THE ANOVA button to compute the classical result of ANOVA and Root Mean Squared Error (RMSE). When the ANOVA result is significant, the Tukey HSD retrospective test can be computed.

6-4 Simple correlation

This procedure calculates correlation between two quantitative variables using Pearson's method (Fig 37).
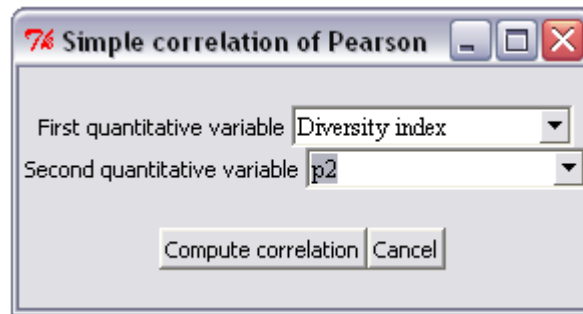
Fig 37: The window to compute simple correlation.

The result gives the Pearson R-squared, its p-value and the equation of the regression. For example, this procedure can be used to test the relationship between the diversity indexes of fingerprint profiles and an environmental parameter.

7- MULTIVARIATE STATISTICS: STRUCTURE MENU

7-1 Non-metric Multidimensional Scaling (nMDS)

Non-metric multidimensional scaling can be based on a proximity matrix computed with any of 13 supported proximity measures, as explained below. Basically, nMDS is a two-dimensional display where each fingerprint profile is represented by a single plot (Cox T.F. & M.A.A., 2001). They are plotted so as to conform as well as possible with the rankings of the proximities between each pair of points. The degree to which the display matches the underlying distances is assessed by using Kruskal stress; a maximum threshold value of 10 indicates little risk of misinterpreting (Clarke & Warwick, 2001).
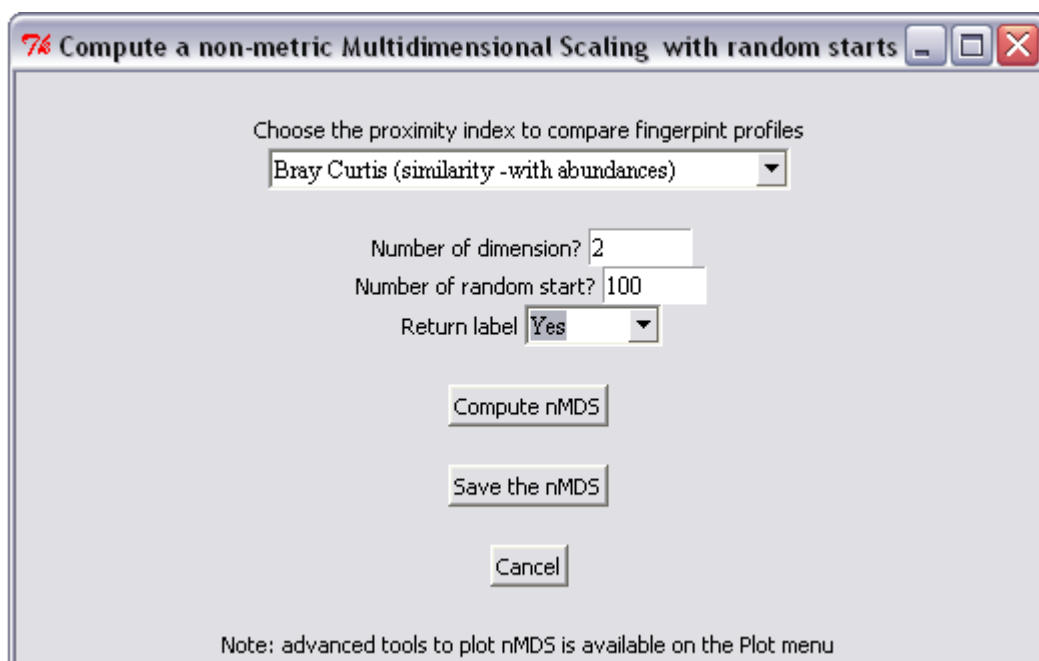

Fig 38: The window to compute nMDS.

The algorithm used in this program allows the proximity index, the number of dimensions and the number of random starts to be chosen (Fig 38). The greater the number of dimensions, the better is the result of nMDS and thus the lower the stress. As the random start procedure provides different plots at each computation, it is recommended to save the nMDS. To view the nMDS in 3 dimensions or to choose the axes to plot or other tools, the nMDS should be saved and then loaded using plot procedures (see sections *5-3 Plot saved nMDS vs PCA in 2 dimensions* and *5-4 Plot saved nMDS vs PCA in 3 dimensions*).

Three kinds of proximity measure are supported (Legendre & Legendre, 1998; Wolda, 1981):

- Distances (Euclidean, Maximum, Manhattan, Canberra, Minkowski). The higher the value, the more different are the compared fingerprint profiles.
- Similarity with abundance (Bray Curtis, Chi –squared, Ruzicka, Roberts). These indexes take into account the relative abundances of each peak. This indexes range from 0 (no proximity) to 1 (the two fingerprint profiles are identical).
- Similarity with presence/absence (Jaccard, Dyce-Sorensen, Ochiai, Steinhaus). These indexes take into account only the presence (value1) or the absence (value 0) on each scan of the fingerprint profiles. This indexes range from 0 (no proximity) to 1 (the two fingerprint profiles are identical). For these indexes, don't forget to transform your fingerprint profiles into presence/absence fingerprint profiles (see section *4-10 Transform profiles into presence/absence profiles*).

7-2 Principal Components Analysis (PCA)

Principal components analysis (PCA) is a procedure for finding hypothetical variables (components) which account for as much of the variance in your multidimensional data as possible. These new variables are linear combinations of the original ones. In this program PCA can be centred and scaled (Fig 39 top).
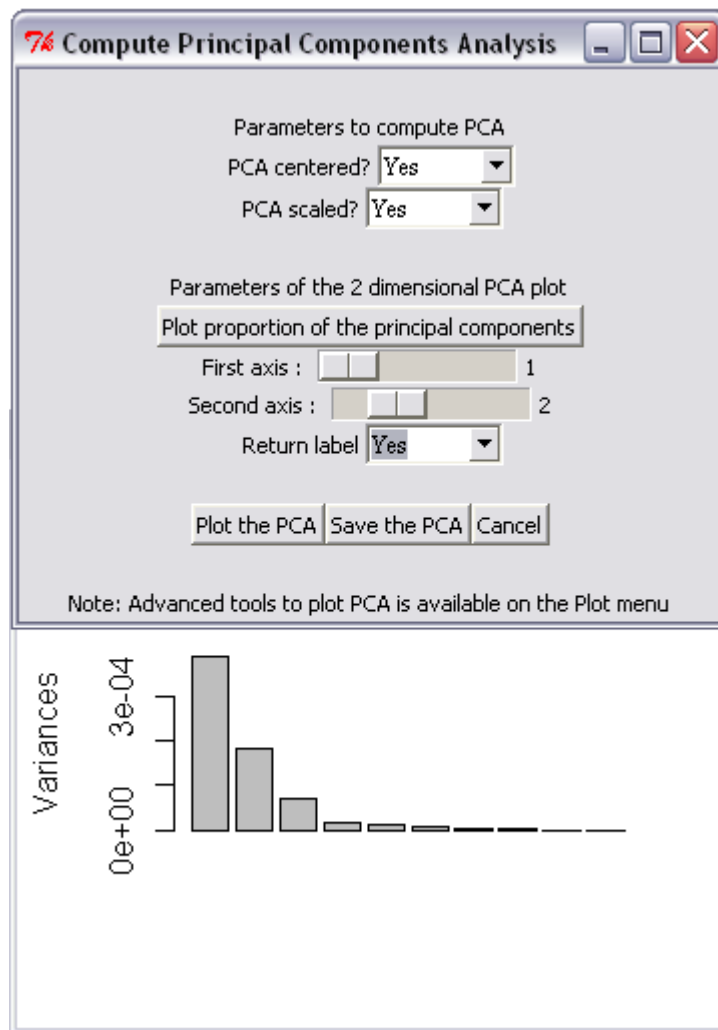
Fig 39: The window to compute PCA.

The PLOT PROPORTION OF THE PRINCIPAL COMPONENTS button which plots the respective contribution of the new components can be helpful to choose the plotting axis (Fig 39, bottom). As for nMDS, PCA can be saved and plotted in 2 or 3 dimensions with advanced features (see sections *5-3 Plot saved nMDS vs PCA in 2 dimensions* and *5-4 Plot saved nMDS vs PCA in 3 dimensions*).

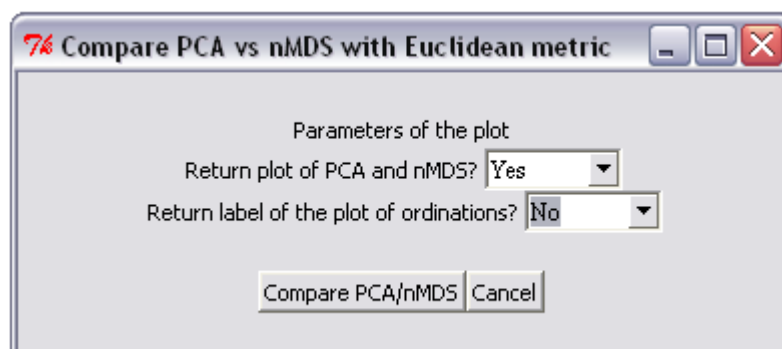## 7-3 Compare PCA vs nMDS

Fig 40: The window for the procedure which compares ordinations.

This procedure compares the nMDS and the PCA ordinations with the Pearson correlation method using a Euclidean metric (Fig 40).

- First calculate the Euclidean distances between fingerprint profiles pairwise (initial distance matrix).
- Next calculate the Euclidean distances between points pairwise, computed by the two ordination methods (the ordination distance matrix).
- Compare the Pearson R-squared between the initial distance matrix and both the ordinations distance matrixes (Fig 41).

Note that the nMDS is calculated with a single random start (see *section 7-1 Non-metric Multidimensional Scaling*).

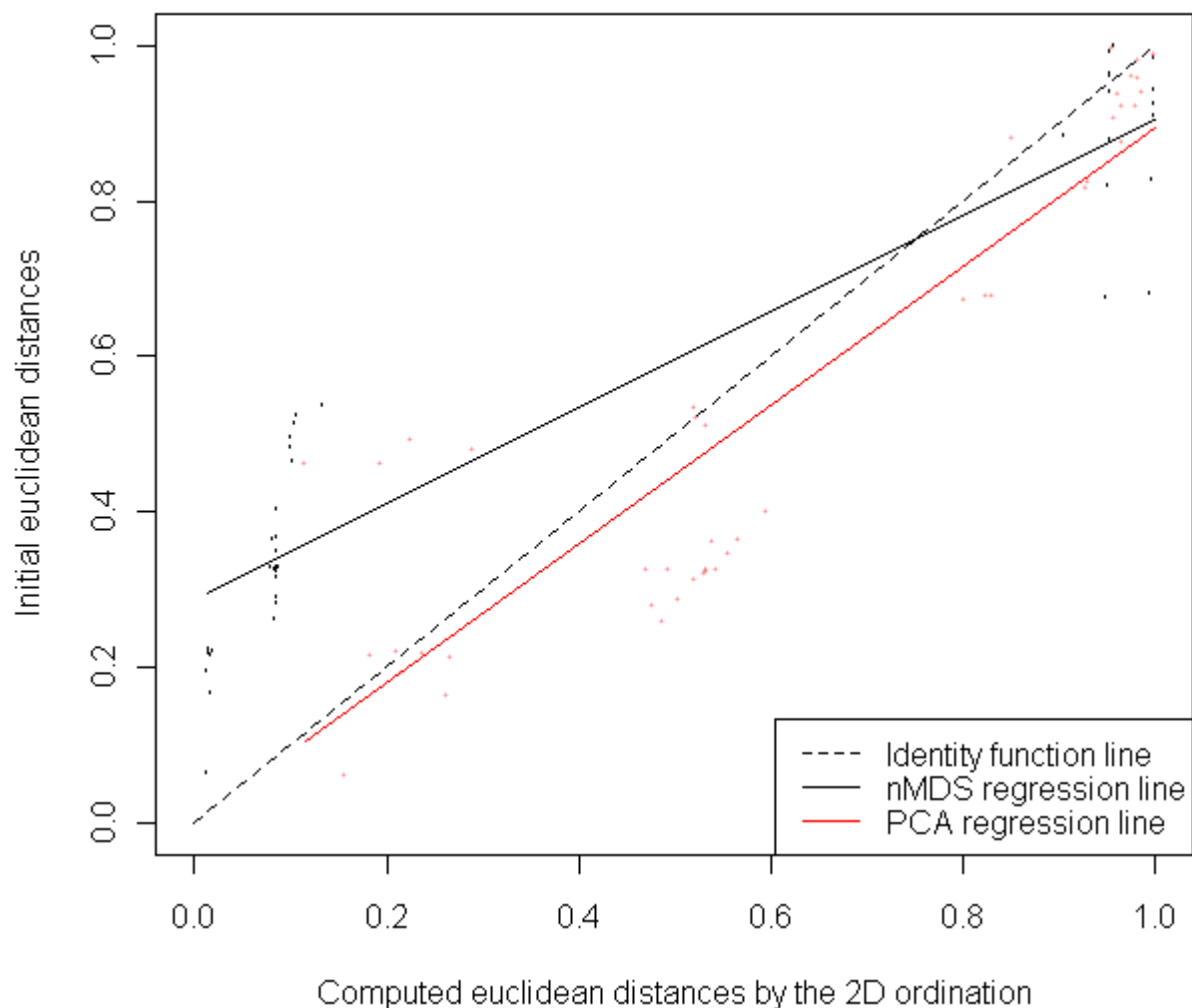## nMDS R^2= 0.87 > PCA R^2 0.8

Fig 41: Resulting plot of the procedure which compares ordinations.

7-4 Hierarchical clustering

The hierarchical clustering algorithm produces a dendrogram with clusters of fingerprint profiles according to their proximities (Fig 42).

Seven different algorithms for plotting the dendrogram are available: ward, single (often named nearest neighbour), complete, average (often named Unweighted Pair-Group Average UPGMA), McQuitty, median and centroid (Gordon, 1999; Murtagh, 1985).

Different proximity measures can be used to compare the fingerprint profiles (see section *7-1 Non-metric Multidimensional Scaling*). However, for Ward's method, Euclidean distance is inherent in the algorithm.
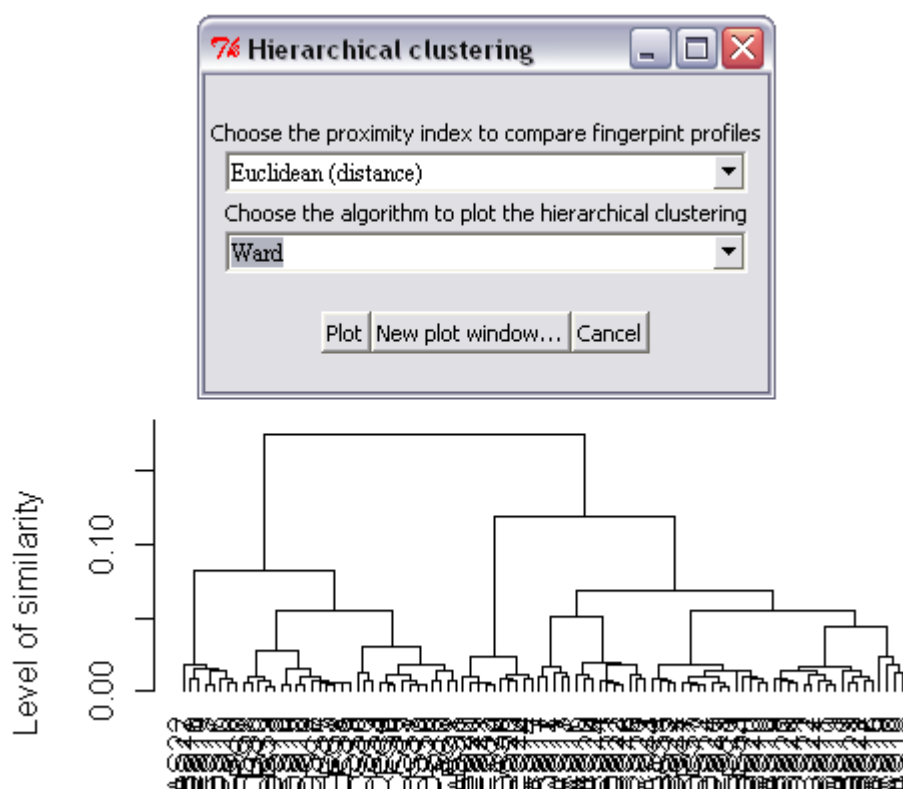

Fig 42: The window to compute hierarchical clustering and the resulting plot.

7-5 Heatmap

A heatmap is a hierarchical clustering with summarized fingerprint profiles added (the signal intensity is proportional to a gradient of colours) to improve the visual interpretation of the plot. (Fig 43). Fingerprint profiles are in rows. Seven and thirteen algorithms can be used for plotting the dendrogram and for calculating proximity measures respectively between fingerprint profiles (see section *7-4 Hierarchical clustering* and section *7-1 Non-metric Multidimensional Scaling* for details on algorithms).
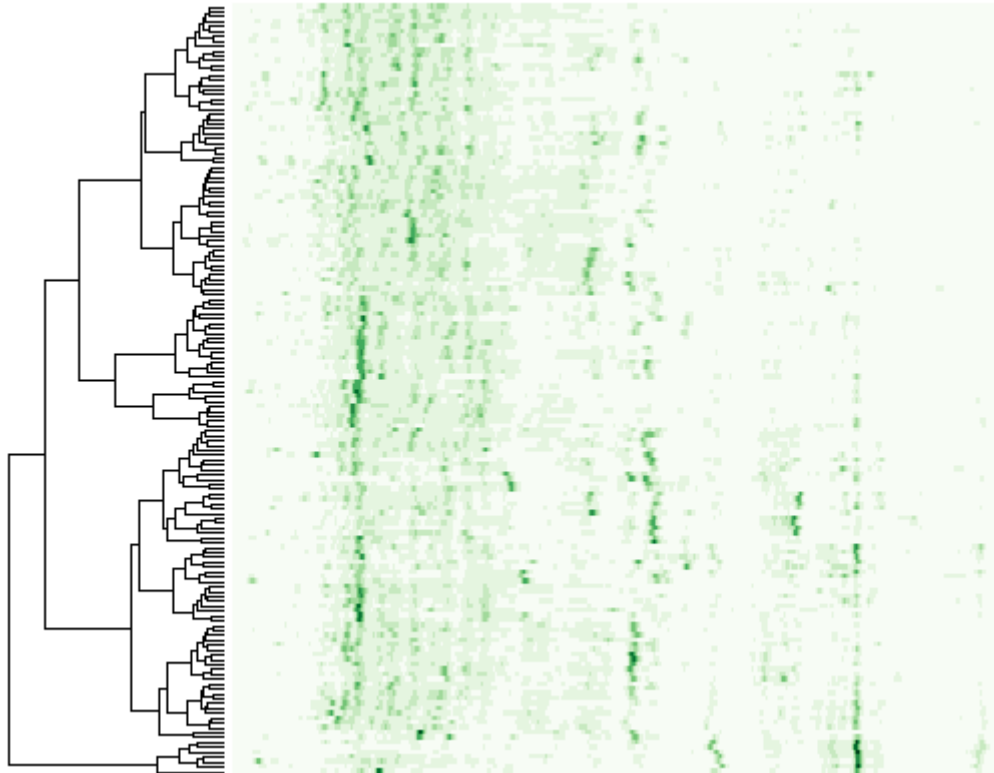
Fig 43: Plot displays by the heatmap procedure.

7-6 Multivariate ANOVA

General linear modelling of fixed-effect models with multiple responses is performed (Fig 44). The procedure calculates 50-50 MANOVA p-values, ordinary univariate p-values and adjusted p-values using rotation testing (Langsrud, 2002; Langsrud et al, 2005; Moen et al, 2005).
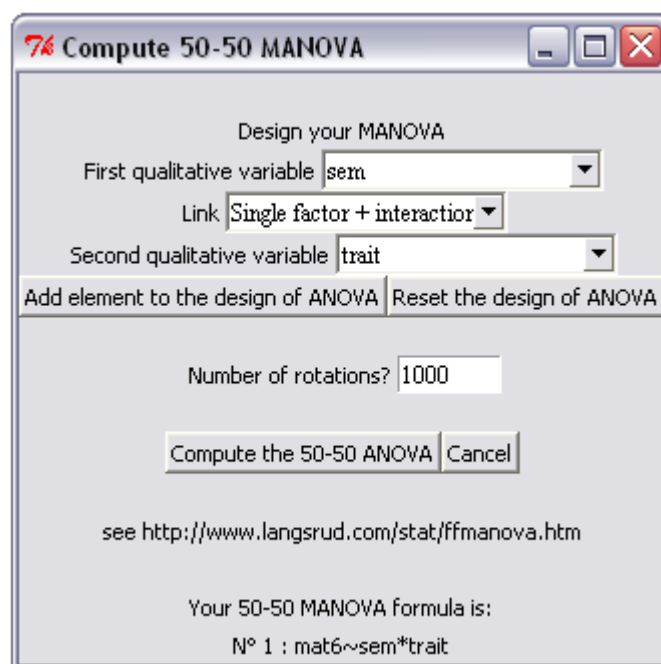
Fig 44: The window to define a design and to compute Multifactor ANOVA.

We advise using at least 1000 rotations to produce an accurate p-value. For help to generate the design see section *6-3 Multifactor ANOVA*.

7-7 Global ANOSIM

ANOSIM (ANalysis Of Similarities) is a non-parametric statistical test of significant difference between two or more groups, based on any proximity measures available (Clarke, 1993) (Fig 45).
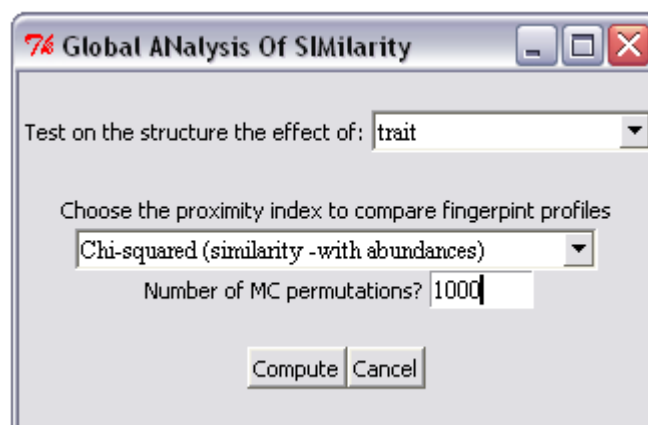

Fig 45: The window to compute global ANOSIM.

Thirteen proximity measures can be chosen (see section *7-1 Non-metric Multidimensional Scaling*). In this procedure, groups are designed according to the level of a factor (qualitative variable). In a rough analogy with ANOVA, the test is

based on comparing distances between groups (rB) with distances within groups (rW) to produce the ANOSIM statistic R:

$$ANOSIM \quad R = \frac{4 \times (rB - rW)}{N \times (N - 1)}$$

This ANOSIM R value ranges from 0 (no difference) to 1 (completely separated groups). The statistical significance of observed ANOSIM R is assessed by Monte Carlo permutations to obtain the empirical distribution of ANOSIM R under the null hypothesis. We advise using at least 1000 permutations to produce an accurate p-value.

7-8 Pairwise ANOSIM

This procedure uses exactly the same algorithm as that described in section *7-7 Global ANOSIM*. It indicates which levels of a factor (qualitative variable) differ from the others (when there is a significant difference according to this qualitative variable using global ANOSIM). This procedure is generally used as a retrospective test after a global ANOSIM or a multivariate ANOVA (see section *7-6 Multivariate ANOVA*).
First select the factor (qualitative variable; Fig 46 top); next choose the two levels to compare (Fig 46 bottom). When selecting "all pairwise ANOSIM" in one or in the two selection boxes of levels (Fig 46 bottom), the returned result is a pairwise ANOSIM (ANOSIM R and p-values) of each pair of levels within the factor.
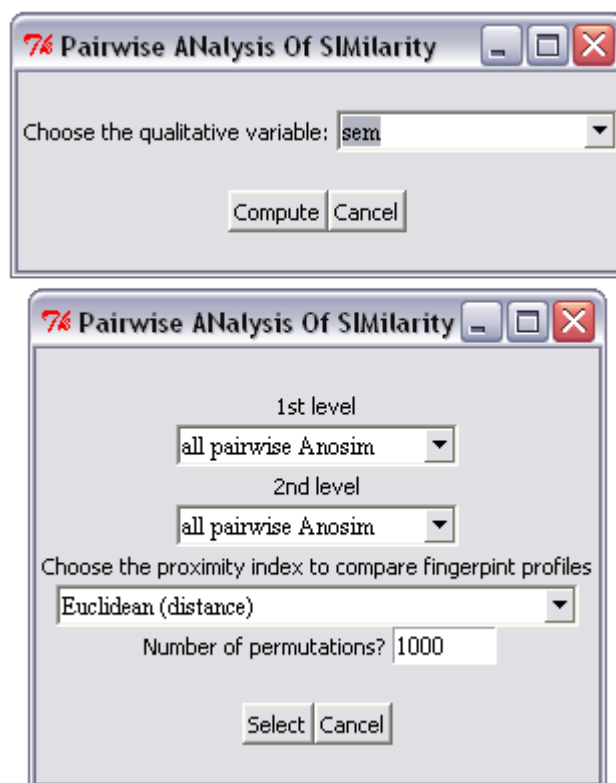

Fig 46: The two consecutive windows to compute pairwise ANOSIM.

7-9 Within-group variability

This procedure tests whether the within-group variability differs significantly for two or more groups of fingerprint profiles (Fig 47).
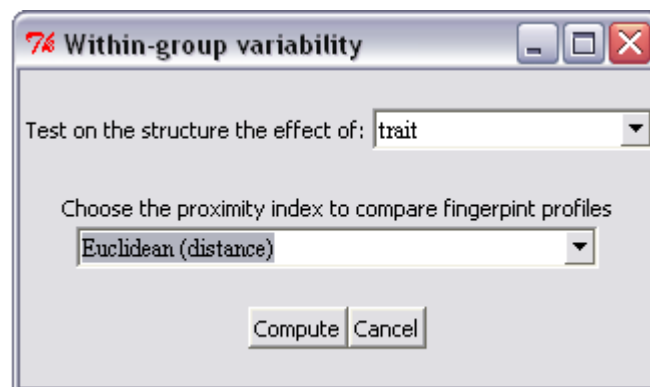

Fig 47: The window to compute the test of the within-group variability.

The groups are defined as levels of a factor (a qualitative variable). The test used by this algorithm consists of a type 2 ANOVA and a TukeyHSD test as a retrospective test. Thirteen proximity measures can be chosen (see section *7-1 Non-metric Multidimensional Scaling*).

This procedure is normally used when fingerprint profiles differ according to the level of a qualitative variable. It shows whether one or more groups of fingerprint profiles (microbial communities) are more or less homogeneous than the other groups (other microbial communities).

7-10 Similarity percentages procedure

SIMPER (SIMilarity PERcentage procedure) is a simple method for assessing which scans are primarily responsible for an observed difference between groups of fingerprint profiles (Clarke, 1993). The overall significance of the difference is often assessed by global ANOSIM (see section *7-7 Global ANOSIM*). In the procedure of this program only the Euclidean distances can be chosen as proximity measure.

Fig 48: The two consecutive windows to compute the SIMPER test.

First select the qualitative variable (Fig 48 top); next choose the two levels to compare (Fig 48 bottom). The output of SIMPER is a list of the scans sorted in decreasing order of contribution to the overall dissimilarity. Their relative and cumulative contributions as percentages are also provided. The procedure also displays a graph with a black curve indicating the relative percentage of contribution of scans (Fig 49). The threshold in per 1000 (Fig 49 bottom) allows scans above this threshold to be coloured red in the resulting plot (Fig 49).

Fig 49: Resulting plot from the SIMPER test.

7-11 Iterative test

An iterative test shows which scans differ significantly at p<0.05 when comparing two groups of fingerprint profiles (two communities). The two groups are defined according to two levels of a qualitative variable. The overall significance of the difference is often assessed by global or pairwise ANOSIM (see section 7-7 Global ANOSIM or 7-8 Pairwise ANOSIM). Thirteen proximity measures can be chosen (see section 7-1 Non-metric Multidimensional Scaling) and three statistical tests: T-test (parametric), Mann Whitney test (non-parametric) and Fisher's exact test (for binary fingerprint profiles).

Fig 50: The two consecutive windows to compute the iterative test on all scans of fingerprint profiles.

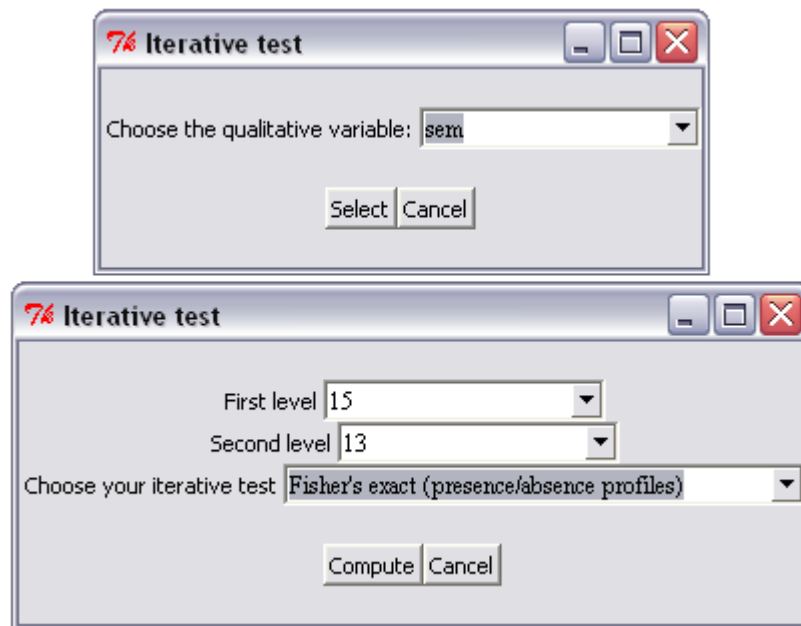First select the qualitative variable (Fig 50 top); then choose the two levels to compare within this qualitative variable (Fig 50 bottom).
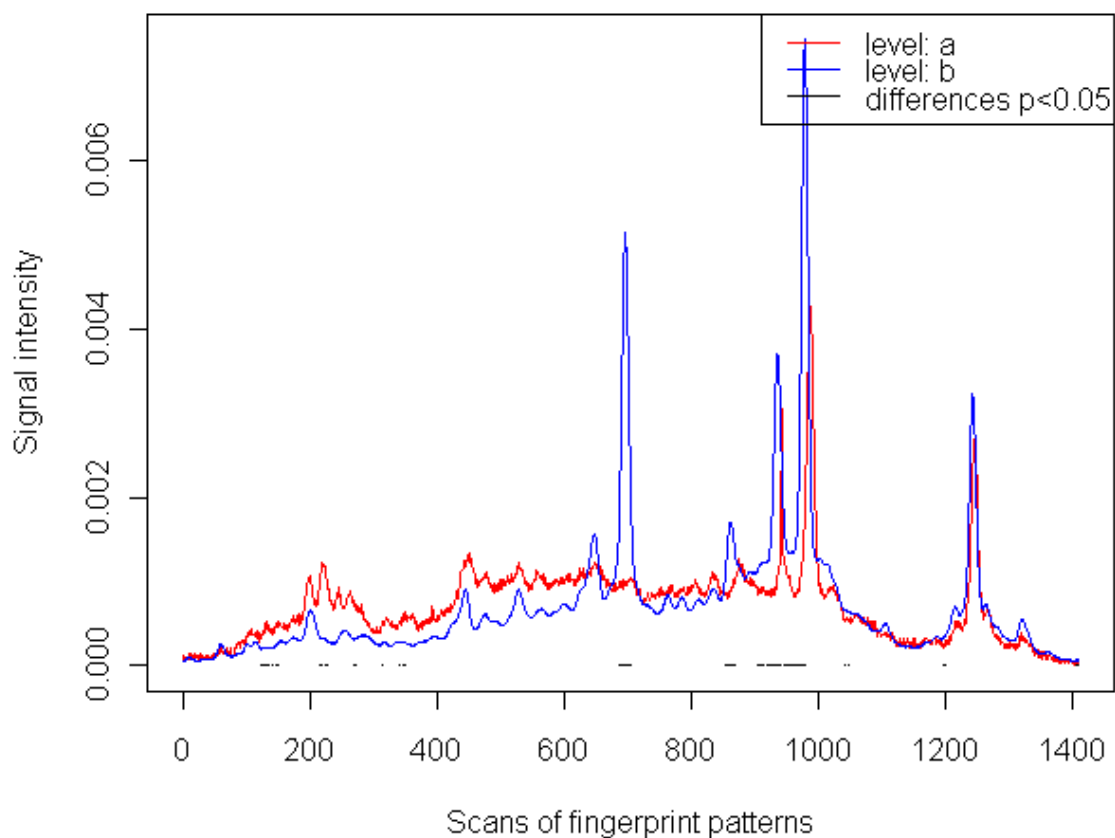

Fig 51: Resulting plot from the iterative test.

The result is a plot with the average fingerprint profiles of each group and the areas of difference as black horizontal lines along the average fingerprint profiles (Fig 51). The percentages as well as the number of significantly different scans are also provided.

7-12 Multivariate correlation

This procedure performs a multivariate correlation test using general linear correlation with a single quantitative variable (Fig 52). P-values, ordinary univariate p-values and adjusted p-values using rotation testing are calculated (Langsrud, 2002; Langsrud et al, 2005; Moen et al, 2005). We advise using at least 1000 rotations to produce accurate p-values.
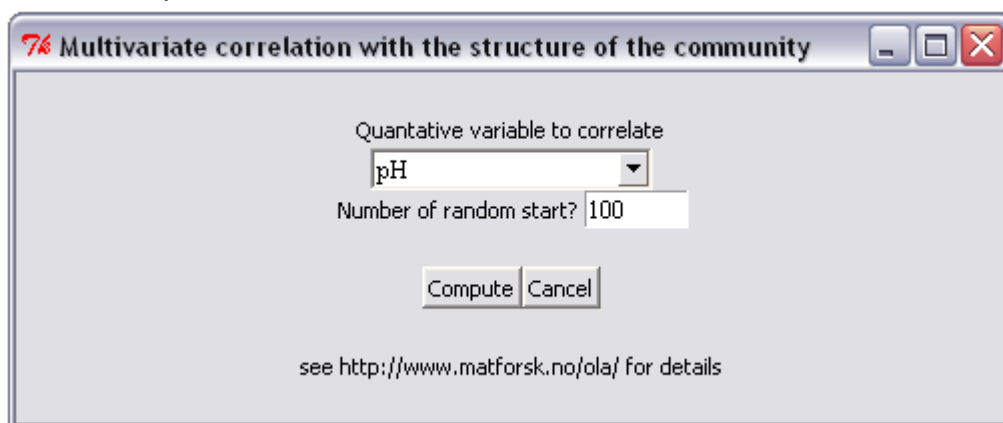

Fig 52: The window to compute the multivariate correlation with the structure of the community.

This procedure is typically used to study the effect of the structure of the microbial community according to environmental parameters such as pH.

8- ADVANCED MODE: INTERNAL OBJECT AND PROCEDURE MANAGEMENT

8-1 Object invoked

Here is a list of the 14 objects used to run a project with STATFINGERPRINTS program:
- filename : this object is the file directory where the Rdata file of your project is located.
- mat6 : this object is a matrix. It contains fingerprint profiles resulting from the last procedure to be executed in the PROFILES PROCESSING menu. All procedures of the menus PLOT, UNIVARIATE STATISTICS: DIVERSITY and MULTIVARIATE STATISTICS: STRUCTURE are executed using this matrix.
- mat : this object is a list. It contains two matrices: the imported raw fingerprint profiles in rows of mat$profile (or mat[[1]]) and their respective internal standard in rows of mat$rox (or mat[[2]]). This object is created when using procedures CONVERT FSA FILES AND IMPORT, IMPORT FINGERPRINT PROFILES IN ASCII FILES or

IMPORT AN ECOLOGICAL TABLE. When using the latter procedure, mat$rox is equal to mat$profil as the internal standards are not loaded.

- mat1 : this object is a matrix containing the align fingerprint profiles resulting from the ALIGN PROFILES ONE BY ONE procedure.
- mat2 : this object is a matrix containing the fingerprint profiles if the DELETE BACKGROUND UNDER PROFILES procedure has been executed; otherwise it is equal to mat1.
- mat3 : this object is a matrix containing fingerprint profiles resulting from the DEFINE A COMMON BASELINE FOR ALL PROFILES procedure.
- mat4 : this object is a matrix containing fingerprint profiles resulting from the DEFINE THE RANGE OF THE PROFILES procedure.
- mat7 : this object is a matrix containing fingerprint profiles resulting from the NORMALIZE AREA UNDER PEAKS procedure.
- mat8 : this object is a matrix containing the fingerprint profiles resulting from the REBUILD PEAKS OF PROFILES PRESENTING DEFAULTS procedure if used; otherwise this object is not defined.
- mat9 : this object is a matrix containing the fingerprint profiles resulting from the TRANSFORM PROFILES INTO PRESENCE/ABSENCE PROFILES procedure if used; otherwise this object is not defined.
- rxref : this object is a vector containing peaks of the reference standard when using procedures DEFINE PEAKS USING YOUR OWN STANDARD or USE PEAKS OF ROX HD400.
- div : this object is a vector containing the last diversity values to be calculated using the procedure COMPUTE DIVERSITY INDEX.
- fact : this object is a matrix containing, in columns, qualitative variables (with levels) loaded when using the procedure IMPORT VARIABLE (QUANTITATIVE OR QUALITATIVE) with selection of qualitative variable in the file type box.
- param : this object is a matrix containing, in columns, quantitative variables loaded when using the procedure IMPORT VARIABLE (QUANTITATIVE OR QUALITATIVE) with selection of quantitative variable in the file type box.
- Alig : this object is a vector containing the names of the fingerprint profiles which were aligned.
- MainMenu : this object is a TclTk object attributed to the main GUI interface.

8-2 Internal procedures

The function to launch the main GUI interface of the program STATFINGERPRINTS is StatFingerprints(). It plots the general GUI interface and has dynamic links to the other procedures.

Most procedures of the STATFINGERPRINTS program are built with 2 functions. The first is the executive function and the second is the GUI function which displays the GUI windows for the executive function. The GUI function is always named with the

name of the executive function plus "GUI" added at the end. Here is a complete list of these functions:

- align(), alignGUI() and alignGUI2() for ALIGN PROFILES ONE BY ONE procedure.
- baseline() and baselineGUI() for DEFINE A COMMON BASELINE FOR ALL PROFILES procedure.
- best.nmds() and best.nmdsGUI() for NON-METRIC MULTIDIMENSIONAL SCALING procedure.
- best.ordination() and best.ordinationGUI() for COMPARISON OF PCA VS NMDS procedure.
- binary() and binaryGUI() for TRANSFORM PROFILES INTO PRESENCE/ABSENCE PROFILES procedure.
- delete.background() and delete.backgroundGUI() for DELETE BACKGROUND UNDER PROFILES procedure.
- diversities() and diversitiesGUI() for COMPUTE DIVERSITY INDEX procedure.
- heatmap() and heatmapGUI() for HEATMAP procedure.
- iterative.test(), iterative.testGUI() and iterative.testGUI1() for ITERATIVE TEST procedure.
- make.peak() and make.peakGUI() for REBUILD PEAKS OF PROFILES PRESENTING DEFAULTS procedure.
- normalisation() and normalisationGUI() for NORMALISE AREA UNDER PROFILES procedure.
- plot3dim() and plot3dimGUI() for PLOT PROFILES IN 3 DIMENSIONS procedure.
- pwanosim() and pwanosimGUI() for PAIRWISE ANOSIM procedure.
- simper(), simperGUI() and simperGUI1() for SIMILARITY PERCENTAGE PROCEDURE procedure.
- trunck() and trunckGUI() for DEFINE THE RANGE OF THE PROFILES procedure.
- within.group.similarity() and within.group.similarityGUI() for WITHIN GROUPS VARIABILITY procedure.

Some procedures of the STATFINGERPRINTS program are only built with a single GUI function as they are available with basic line commands in the R prompt:

- addGUI() for ADD PROFILES TO THE PROJECT procedure.
- anovaGUI() for MULTIFACTOR ANOVA procedure.
- askquitGUI() for QUIT R procedure.
- change.nameGUI() for CHANGE NAMES OF PROFILES procedure.
- convertGUI() for CONVERT FSA FILES AND IMPORT procedure.
- correlationGUI() for SIMPLE CORRELATION procedure.
- deleteGUI() for DELETE PROFILES TO THE PROJECT procedure.
- dendoGUI() for HIERARCHICAL CLUSTERING procedure.
- des.univGUI() for DESCRIPTIVE STATISTICS procedure.
- ffcorGUI() for MULTIVARIATE CORRELATION procedure.
- globalanosimGUI() for GLOBAL ANOSIM procedure.
- importGUI1() for IMPORT FINGERPRINT PROFILES IN ASCII FILES procedure.

- importGUI2() for IMPORT VARIABLES (QUANTITATIVE/QUALITATIVE) procedure.
- importGUI3() for IMPORT FINGERPRINT PROFILES IN AN ECOLOGICAL TABLE procedure.
- loadGUI() for LOAD PROJECT procedure.
- pcaGUI() for PRINCIPAL COMPONENTS ANALYSIS procedure.
- permanovaGUI() for MULTIVARIATE ANOVA procedure.
- plot2dGUI() for PLOT PROFILES IN 2 DIMENSIONS procedure.
- plotord3dGUI() for PLOT SAVED NMDS VS PCA IN 3 DIMENSIONS procedure.
- plotordGUI() and nmds.contour() for PLOT SAVED NMDS VS PCA IN 2 DIMENSIONS procedure.
- roxnewGUI() for DEFINE PEAKS USING YOUR OWN REFERENCE STANDARD procedure.
- saveasGUI() for SAVE PROJECT AS… procedure.
- saveGUI() for SAVE PROJECT procedure.

Some procedures of the STATFINGERPRINTS program are only built with a single executive function as they do not need a GUI interface or are used in routine for another executive function:

- clearconsole() for clearing the output and warning consoles of the main GUI interface.
- newdist() for computing proximity measures between fingerprint profiles when using functions: best.nmds(), dendoGUI(), heatmap(), globalanosimGUI(), pwanosim() and simper().
- Newrdata() for creating empty objects needed to create a new STATFINGERPRINTS projects.
- peakparameters() for detecting peaks when using functions: binary() and diversities().
- rollball() for defining the background under peaks when using functions: binary(), diversities() and delete.background().
- roxdefault() for use PEAKS OF ROX HD400 procedure.
- Oppdf() for opening PDF document of the HELP menu.

CONCLUSION

STATFINGERPRINTS is helpful for microbial ecologists to analyse fingerprint profiles. Using a single package, it allows the output of the sequencer to be analysed statistically. It provides numerous procedures for processing raw fingerprint profiles, calculating diversity indexes and performing statistical analyses. The multivariate statistics part is more developed than the univariate one for the majority of the multivariate tests used. Furthermore it is a free program, easy to use, with a graphical user interface and running on numerous platforms (UNIX, Windows, MacOS).

ACKNOWLEDGEMENT

REFERENCES

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Austral Ecol 18(1): 117-143

Clarke KR, Warwick RM (2001) Change in marine communities: an approach to statistical analysis and interpretation, 2 edn. Plymouth, UK.

Cox T.F., M.A.A. C (2001) Multidimensional Scaling.

Forsythe GE, Malcolm MA, Moler CB (1977) Computer Methods for Mathematical Computations.

Gordon AD (1999) Classification, second edn. London.

Langsrud (2002) 50-50 multivariate analysis of variance for collinear responses. The Statistician 51: 305-317

Langsrud, Oslash, yvind (2005) Rotation tests. Statistics and Computing 15: 53-60

Legendre P, Legendre L (1998) Numerical ecology., 2 edn. Amsterdam: Elsevier.

Magurran AE (2004) Measuring biological diversity., Oxford: Blackwell publishing.

Moen B, Oust A, Langsrud O, Dorrell N, Marsden GL, Hinds J, Kohler A, Wren BW, Rudi K (2005) Explorative Multifactor Approach for Investigating Global Survival Mechanisms of Campylobacter jejuni under Environmental Conditions. Appl Environ Microbiol 71(4): 2086-2094

Murtagh F (1985) Multidimensional Clustering Algorithms. In Compstat. Wuerzburg

R development Core Team (2008) R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing.

Wolda H (1981) Similarity indices, sample size and diversity. Oecologia 50: 296-302