Molecular Ecology Notes (2007) 7, 767–770 doi: 10.1111/j.1471-8286.2007.01882.x

PROGRAM NOTE

SAFUM: statistical analysis of SSCP fingerprints using PCA projections, dendrograms and diversity estimators

O. ZEMB,*† B. HAEGEMAN,*‡ J. P. DELGENES,* P. LEBARON† and J. J. GODON*
*INRA, UR050, Laboratoire de Biotechnologie de l'Environnement, Avenue des Etangs, Narbonne F-11100, France, †Université Pierre et Marie Curie, Paris 6, Paris F-75005, France; Institut National des Sciences de l'University, CNRS, UMR7621, BP 44, Banyuls-sur-Mer F-66650, France, ‡INRA-INRIA research team MERE, UMR Analyse des Systèmes et Biométrie, 2 place Pierre Viala, Montpellier F-34060, France

Abstract

The program SAFUM provides a smart interface to import, visualize and compare fingerprinting profiles, especially on capillary electrophoresis single strand conformation polymorphism data, in conjunction with basic statistical analysis tools. It includes principal component analysis, two- or three-dimensional representations, dendrograms based on Euclidean distance, and easily exportable files for subsequent applications. SAFUM is useful for the analysis of spatial or temporal sequences of microbial community fingerprints obtained with an ABI prism sequencer.

Keywords: capillary electrophoresis, fingerprint, microbial, SSCP, statistical analysis, diversity

Received 11 March 2007; revision accepted 27 May 2007

Molecular methods based on DNA extraction and polymerase chain reaction (PCR) amplification stand as key steps in molecular microbial ecology studies (Hewson & Fuhrman 2006). The PCR product can be analysed by sequencing randomly picked fragments (Lozupone *et al*. 2006), or by carrying out a differential migration of PCR fragments generating a fingerprint of the community. Capillary electrophoresis single strand conformation polymorphism (CE-SSCP) is an example of the latter (Delbes *et al*. 2000). The electrophoretic separation is based on differences in secondary structure of single-stranded nucleic acids and corresponding differences in their mobility through a gel. The fragments are detected at the end of the capillary, thus providing a retention profile. An internal standard is analysed simultaneously so that these profiles can be compared quantitatively. Even if the technique is not so difficult (Sunnucks *et al*. 2000), the analysis might be a bit harder, since different authors use different ad hoc methods that might not be easily transposable. As many authors use fingerprints (Delbes *et al*. 2001; Godon *et al*. 2001; Dabert *et al*. 2005; Callon *et al*. 2006; Hori *et al*. 2006; Peu *et al*. 2006), analyses are diverse, making comparison of data sometimes difficult. A frequent use of these fingerprints makes their systematic manipulation critical. The aim of the software is to allow easy and transparent comparison that is needed to study population reproducibility (Collins *et al*. 2006; Hoshino *et al*. 2006; Ranjard *et al*. 2006) and activity (Weinbauer *et al*. 2002; Hewson *et al*. 2006; Wertz *et al*. 2006). In order to compare profiles, an automated analysis is convenient and objective. Several commercial packages are available, but they do not allow data manipulation beyond preprogrammed algorithms, and they are often expensive and not transparent enough for a satisfactory use in microbial ecology. For example, it was recently demonstrated that the area under the peaks contains information (Loisel *et al*. 2006). Commercial software packages have difficulties to deal with raw data taking this area into account. An exception is the MEDIMECO MATLAB (MathWorks) script for fingerprint analysis developed by Harmand *et al*. (2006) but its parameterization and use require numerous manipulations and extensive MATLAB knowledge.

In this note, we present the open-source program SAFUM, also written in MATLAB but with a graphical user interface. SAFUM can be downloaded on http://www.montpellier.inra.fr/narbonne/anglais/researchunits/microbialecology.html. No programming knowledge is required but the purchase of MATLAB is necessary. Most of the features of MEDIMECO are available and some extra

Correspondence: Olivier Zemb, Fax: (33) 4 68 42 51 60 (outside France); E-mail: zemb.olivier@gmail.com

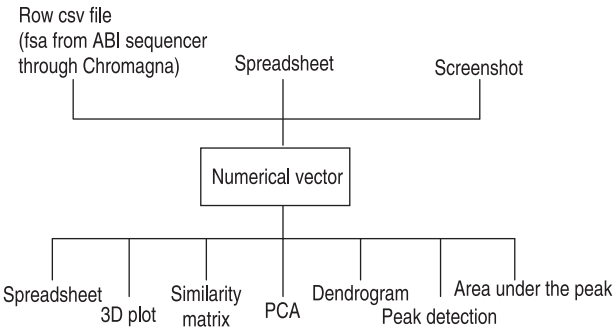

Fig. 1 Block representation of SAFUM structure. The input files can be spreadsheets containing unprocessed signals extracted from FSA files with CHROMAGNA, spreadsheets with each column containing the fingerprint of one sample, or screenshots. The process in SAFUM allows performing peak detection, principal component analysis and dendrograms. Fingerprints, peak abscissa, peak area and area under the peaks in spreadsheets can be exported in spreadsheets.

possibilities are accessible. We demonstrate some possibilities of SAFUM on a typical example, by analysing fingerprints of a semicontinuous anaerobic reactor fed with a mixture of wine and milk operated under stable conditions until day 89 when an organic overload occurred. The organic overload resulted in an acidification of the reactor. After 3 days, the pH was set back to normal.

The first step consists of importing the profiles from an ABI prism sequencer. SAFUM offers two ways of importing data (Fig. 1): GIF graphical data files from a screenshot of the built-in software of the sequencer, or text files containing the sample and the internal standard data in comma-separated value (CSV) format. When using an ABI prism sequencer, these CSV files are obtained from the program CHROMAGNA, that was developed by M. J. Miller at the US National Institute of Health in order to read the files generated by ABI GENSCAN or GENEMAPPER (Fekete *et al*. 2003). A folder can be chosen everywhere on the computer and all the CSV files will be automatically selected. SAFUM also provides an alignment algorithm when CSV raw data are used as input files. These files should contain both sample and internal standard information to be aligned. A reference profile using the ROX400HD (Applied Biosystems) is included in SAFUM to guarantee identical alignment for different sets of samples. This reference profile can easily be modified by the user if other ladders are used, by replacing the 'roxref.csv' file in the SAFUM folder.

The total area generated by the signal can be normalized to one so that relative abundance of each peak can be compared (Brown *et al*. 2005). All normalized spectra can be stacked together allowing immediate comparison. Figure 2a shows a few peaks that migrate at the same position in several samples, and rise or decrease as a function of time. For a set of normalized profiles, the Euclidean distance between two spectra can be computed. This normalized distance is computed taking into account all the data (scans in case of CSV input or pixels in case of image input) and not only peak areas as most software packages do. The more two profiles mismatch, the bigger this distance. Based on the Euclidean distances, SAFUM computes corresponding dendrograms using different methods (for more details see Legendre & Legendre 1998). SAFUM also displays the divergence rate as a function of time in case of temporal sequences, or as a function of location in case of spatial sequences. In Fig. 2b, the last two spectra are different, due to the overloading event happening on day 90 (Zemb, PhD thesis, 2007).

Principal component analysis (PCA) summarizes large data sets into a small number of vectors gathering most of the variability, and allows to produce representative plots

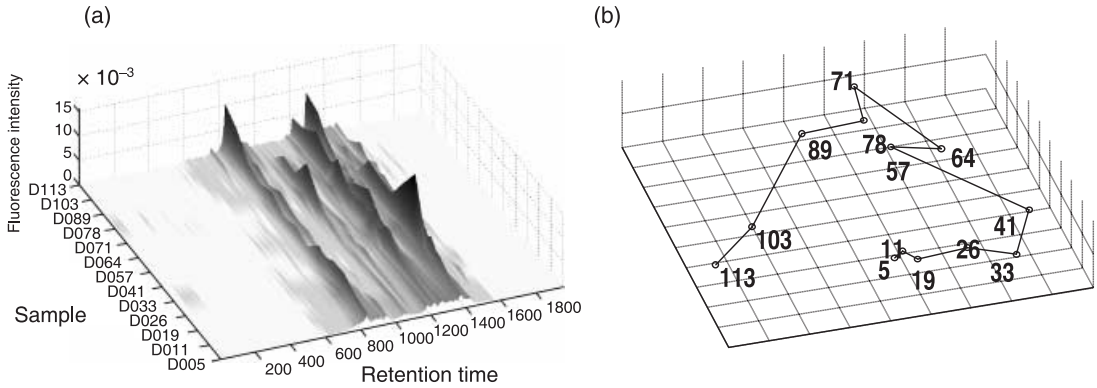

Fig. 2 Single strand conformation polymorphism spectra from an anaerobic digester under stable operation. (a) Three-dimensional view. The retention time (*X*-axis) varies between 0 and 2000. The names of the samples contain the number of days after the beginning of the experiment. The area of each fingerprint was normalized to one. (b) Principal component analysis (PCA) representation. The first three principal components explain 74% of the variability (42% for PCA1, 21% for PCA2 and 11% for PCA3). Numbers indicate the number of days after the inoculation.

(Ranjard *et al*. 2006). In our case, each SSCP profile is a vector with a large number of variables (typically 1000). PCA looks for linear combinations of these variables, which best represent the variance present in the data. SAFUM displays the data on the first two or three principal components (called PCA1, PCA2 and PCA3). Simultaneously, the variance present in the projection of the data in these principal components is compared to the total data variance. In this way, the user can judge whether the reduced representation is meaningful. In our example, 74% of the data variance is represented on the first three principal components. Distances between the projected samples were computed, using thus only three-fourths of the genetic changes. Nevertheless, Fig. 2a shows that the bacterial community of the anaerobic digester evolves in a continuous way, as reported in previous studies (Fernandez *et al*. 1999; Zumstein *et al*. 2000). Interestingly, the distance between fingerprints after the perturbation (of day 103 and day 113) is much larger than during the stable period.

As described previously (Loisel *et al*. 2006), the area under the peaks contains some information. To our knowledge, SAFUM is the first software that allows access to this information, which is mostly removed during the data recovery. The peak areas and area under the peaks can be used for comparison of samples in terms of diversity or species abundance distribution (Hamelin, personal communication).

In conclusion, SAFUM provides a solution for importing data from capillary sequencers to facilitate specific use of fingerprints in microbial ecology. Once data are imported, all steps are transparent and statistical tools like PCA and dendrograms are available through a user–friendly interface. All results are easily exportable in spreadsheet formats.

Acknowledgements

We thank Dimitri Vanpetghem for his help about background removal algorithms. We thank Laurent Lardon and Jérôme Harmand for their help in the MATLAB work, Jérôme Hamelin for very helpful discussion. We thank Jérôme Hamelin, Nyree West, Jeff Ghiglione, Patrice Rey and Valérie Bru for their suggestions after their use of SAFUM.

References

Brown MV, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environmental Microbiology*, **7**, 1466–1479.

Callon C, Delbes C, Duthoit F, Montel MC (2006) Application of SSCP–PCR fingerprinting to profile the yeast community in raw milk Salers cheeses. *Systematic and Applied Microbiology*, **29**, 172–180.

Collins G, Mahony T, O'Flaherty V (2006) Stability and reproducibility of low-temperature anaerobic biological wastewater treatment. *Fems Microbiology Ecology*, **55**, 449–458.

Dabert P, Delgenes JP, Godon JJ (2005) Monitoring the impact of bioaugmentation on the start up of biological phosphorus removal in a laboratory scale activated sludge ecosystem. *Applied Microbiology and Biotechnology*, **66**, 575–588.

Delbes C, Moletta R, Godon JJ (2000) Monitoring of activity dynamics of an anaerobic digester bacterial community using 16S rRNA polymerase chain reaction—single-strand conformation polymorphism analysis. *Environmental Microbiology*, **2**, 506–515.

Delbes C, Moletta R, Godon J (2001) Bacterial and archaeal 16S rDNA and 16S rRNA dynamics during an acetate crisis in an anaerobic digestor ecosystem. *FEMS microbiology ecology*, **35**, 19–26.

Fekete RA, Miller MJ, Chattoraj DK (2003) Fluorescently labeled oligonucleotide extension: a rapid and quantitative protocol for primer extension. *BioTechniques*, **35**, 90–94.

Fernandez A, Huang SY, Seston S *et al*. (1999) How stable is stable? Function versus community composition. *Applied and Environmental Microbiology*, **65**, 3697–3704.

Godon JJ, Duthoit F, Delbes C, Millet L, Montel MC (2001) Use of molecular fingerprint for the study of complex microbial ecosystem. Application to AOC Salers cheese. *Lait*, **81**, 257–262.

Harmand J, Paulou L, Desmoutiers J *et al*. (2006) The microbial signature of drinking waters: myth or reality? *Water Science and Technology*, **53**, 259–266.

Hewson I, Fuhrman JA (2006) Improved strategy for comparing microbial assemblage fingerprints. *Microbial Ecology*, **51**, 147–153.

Hewson I, Steele JA, Capone DG, Fuhrman JA (2006) Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Marine Ecology — Progress Series*, **311**, 67–77.

Hori T, Haruta S, Ueno Y, Ishii M, Igarashi Y (2006) Dynamic transition of a methanogenic population in response to the concentration of volatile fatty acids in a thermophilic anaerobic digester. *Applied and Environmental Microbiology*, **72**, 1623–1630.

Hoshino T, Terahara T, Yamada K *et al*. (2006) Long-term monitoring of the succession of a microbial community in activated sludge from a circulation flush toilet as a closed system. *Fems Microbiology Ecology*, **55**, 459–470.

Legendre P, Legendre L (1998) *Numerical ecology*, 2nd English edn. Elsevier, Amsterdam, The Netherlands.

Loisel P, Harmand J, Zemb O *et al*. (2006) Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. *Environmental Microbiology*, **8**, 720–731.

Lozupone C, Hamady M, Knight R (2006) UNIFRAC — an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.

Peu P, Brugere H, Pourcher AM *et al*. (2006) Dynamics of a pig slurry microbial community during anaerobic storage and management. *Applied and Environmental Microbiology*, **72**, 3578–3585.

Ranjard L, Lignier L, Chaussod R (2006) Cumulative effects of short-term polymetal contamination on soil bacterial community structure. *Applied and Environmental Microbiology*, **72**, 1684–1687.

Sunnucks P, Wilson AC, Beheregaray LB *et al*. (2000) SSCP is not so difficult: the application and utility of single-stranded

conformation polymorphism in evolutionary biology and molecular ecology. *Molecular Ecology*, **9**, 1699–1710.

Weinbauer MG, Fritz I, Wenderoth DF, Hofle MG (2002) Simultaneous extraction from bacterioplankton of total RNA and DNA suitable for quantitative structure and function analyses. *Applied and Environmental Microbiology*, **68**, 1082–1087.

Wertz S, Degrange V, Prosser JI *et al*. (2006) Maintenance of soil functioning following erosion of microbial diversity. *Environmental Microbiology*, **8**, 2162–2169.

Zumstein E, Moletta R, Godon JJ (2000) Examination of two years of community dynamics in an anaerobic bioreactor using fluorescence polymerase chain reaction (PCR) single-strand conformation polymorphism analysis. *Environmental Microbiology*, **2**, 69–78.