Journal of Microbiological Methods 79 (2009) 89–95

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Functional principal component data analysis: A new method for analysing microbial community fingerprints

Janine B. Illian *, James I. Prosser, Kate L. Baker, J. Ignacio Rangel-Castro [1]

Institute of Biological and Environmental Sciences, University of Aberdeen, Cruickshank Building, St. Machar Drive, Aberdeen, AB24 3UU, UK

ARTICLE INFO

Article history:
Received 16 March 2009
Received in revised form 14 August 2009
Accepted 18 August 2009
Available online 23 August 2009

Keywords:
Functional analysis
DGGE
T-RFLP
Microbial diversity
Community structure

ABSTRACT

A common approach to molecular characterisation of microbial communities in natural environments is the amplification of small subunit (SSU) rRNA genes or genes encoding enzymes essential for a particular ecosystem function. A range of 'fingerprinting' techniques are available for the analysis of amplification products of both types of gene enabling quantitative or semi-quantitative analysis of relative abundances of different community members, and facilitating analysis of communities from large numbers of samples, including replicates. Statistical models that have been applied in this context suffer from a number of unavoidable limitations, including lack of distinction between closely adjacent bands or peaks, particularly when these differ significantly in intensity or size. Current approaches to the analysis of banding structures derived from gels are typically based on standard multivariate analysis methods such as principal component analysis (PCA) which do not consider structure of DGGE gels but treat the intensity of each band as independent from the other bands, ignoring local neighbourhood structures. This paper assesses whether a new statistical analytical technique, based on functional data analysis (FDA) methods, improves the discriminatory ability of molecular fingerprinting techniques. The approach regards band intensities as a mathematical function of the location on the gel and explicitly includes neighbourhood structure in the analysis. A simulation study clearly reveals the weaknesses of the standard PCA approach as opposed to the FDA approach, which is then used to analyse experimental DGGE data.

1. Introduction

Microbial communities in natural environments are now routinely characterised using molecular techniques, which avoid the problems of bias and selectivity associated with traditional cultivation-based techniques and enable detection and phylogenetic analysis of organisms which cannot be, or have not been, cultivated in the laboratory (Cardenas and Tiedje, 2008). A common approach to molecular characterisation is the amplification of small subunit (SSU) rRNA genes or genes encoding enzymes essential for a particular ecosystem function (Prosser and Embley, 2002). The former approach facilitates taxonomic description of the organisms present, while the latter provides information on the composition of communities carrying out a particular process. A range of 'fingerprinting' techniques are available for analysis of amplification products of both types of gene, including denaturing gradient gel electrophoresis (DGGE) (Muyzer et al., 1993; Muyzer and Smalla, 1998; Torsvik and Øvreås, 2002), temperature gradient gel electrophoresis (TGGE) (Muyzer, 1999), terminal restriction fragment length polymorphism analysis (T-RFLP) (Liu et al., 1997; Smalla et al., 2007) and single strand conformation polymorphism analysis (SSCP) (Sunnucks et al., 2000; Smalla et al., 2007). These techniques allow more rapid analysis than cloning and sequencing of amplification products. They also enable quantitative or semi-quantitative analysis of relative abundances of different community members, and facilitate analysis of communities from large numbers of samples, including replicates.

A number of methods have been used for quantitative analysis of fingerprint patterns and comparison of communities (Ramette, 2009). Statistical models applied to study plant and animal distributions in population ecology are gaining relevance among microbial ecologists, but the adaptation of such methods is still under development and these are mostly used for exploratory analysis of molecular data (Ramette, 2007). All methods suffer from a number of unavoidable limitations including bias resulting from samples run on different gels or machines; PCR bias or low quality amplification products; differences in detection limits, which affect the interpretation of "false" and "true" bands or peaks; and lack of distinction between closely adjacent bands or peaks, particularly when these differ significantly in intensity or size (Nakatsu, 2007; Thies, 2007).

The aim of this study was to determine whether a new statistical analytical technique, based on functional data analysis (FDA) methods,

* Corresponding author. Present address: School of Mathematics and Statistics, CREEM, the Observatory, Buchanan Gardens, University of St Andrews, St Andrews, Fife KY16 9LZ, UK. Tel.: +44 1334 461803; fax: +44 1334 461 800.
E-mail address: janine@mcs.st-and.ac.uk (J.B. Illian).
[1] Present address: Department of Biomedicine, School of Health and Medical Sciences. University of Örebro, 701 85 Örebro, Sweden.

0167-7012/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2009.08.010

could improve the ability of molecular fingerprinting techniques to compare DGGE profiles. The approach is appropriate for functional data, where observations are mathematical functions. In the given context the data are band intensities as a function of location on the gel. To test the suitability of this approach, FDA was used to determine and compare the community compositions of several simulated communities, as characterised by DGGE, and then used to analyse experimental DGGE data.

1.1. Rationale

In the absence of noise and with no limitation on replication, even the simplest of statistical methods is capable of distinguishing between very similar communities, as measured by DGGE profiles. However, variability within samples and variability associated with experimental processing require analysis of several replicates, particularly in cases of higher similarity between communities, to estimate experimental variability between replicates. In addition, the banding profiles or fingerprints obtained from DGGE analysis may exhibit only a few bands in common among groups of similar samples, and distribution of bands may vary across a profile. As a consequence, most applications require careful choice of the appropriate analysis method for specific types of data sets. In the context of DGGE analysis, a number of common problems occur that may have an impact on the quality of the analysis, and several sources of noise will influence the quality of the data. For example, microbial community composition will vary between samples, even if they have been taken at the same location, and technical noise resulting from the gel analysis itself may have an impact on the outcome, independently of the quality of the original sample. This type of noise will increase the variation within the groups. In contrast, the variation between the groups might be rather low in many studies. It is expected that the microbial structure of, say, groups of soil samples collected on an arable field and in a desert will differ substantially. However, a great number of studies deal with situations where the samples in the different groups are very similar; in other words there are large numbers of bands in common (Fromin et al., 2002; Nakatsu, 2007). A further technical problem is the difficulty in locating specific bands in complex profiles of different communities and consequent incorrect alignment and identification. Similar migration patterns, e.g. bands located near each other on a gel even when they are not even phylogenetically closely related, make the data analysis difficult. Current approaches to the analysis of banding structures derived from gels are typically based on standard multivariate analysis methods such as principal component analysis (PCA) (Ramette, 2007). However, these do not consider the structure of DGGE gels but treat the intensity of each band as independent from the other bands, ignoring local neighbourhood structures.

In order to take the neighbourhood structure in the gel into account, information on the location of each band along the gel must be included in the analysis. This may be achieved using FDA methods, in particular functional principal component analysis (FPCA). This method considers the strength of the expression (e.g. relative intensity of a DGGE band) relative to the location of the bands on the gel, i.e. it treats band intensity as a function of location on the gel. Fig. 1 shows a schematic of a simulated banding pattern (a) and the associated function (b).

This paper compares the proposed approach to a standard PCA approach in a detailed simulation study and applies it to DGGE profiles derived from soil samples.

2. Methods

2.1. Functional data analysis

Functional data analysis methods comprise a group of statistical methods that handle data sets of functional data, i.e. data that are mathematical functions rather than single values (Ramsay and Silverman, 2005). Typically, these are functions of time, but functions of distance have also been considered (Illian et al., 2006). The applications of all these methods assess values that are theoretically continuous functions but that might have been measured at (many) discrete time points. Examples include growth curves or heart rates over time or spatial autocorrelation as a function of distance. Many standard statistical methods have been generalised to be applicable to functional data and Ramsay and Silverman (2005) provide a detailed overview of functional data analysis methods. Functional data analysis methods

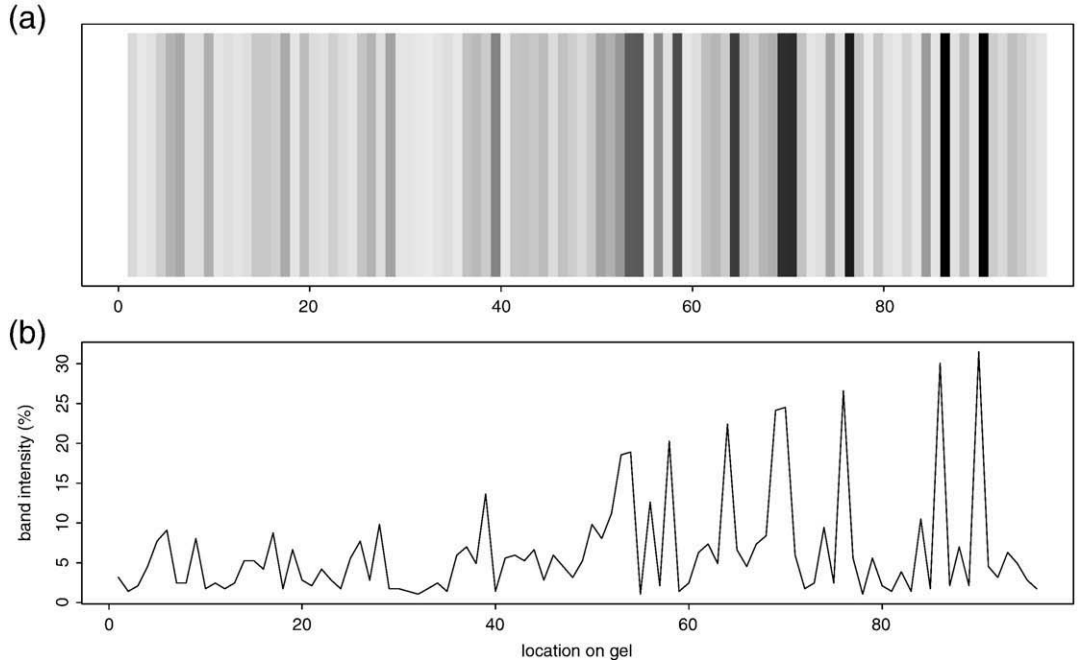

Fig. 1. Banding pattern of a simulated artificial community (a) and the associated function (b).

have been implemented in the commercial statistical software Splus as an add-on library (Clarkson et al., 2005) and the open-source software R (R-development Team, 2005).

Here, functional data analysis methods are applied to the analysis of DGGE profiles. These consist of discrete bands of different intensity and distributed along a gel. FDA treats such DGGE intensity profiles as a continuous function of distance along the gel. More specifically, FPCA is applied to identify groups of similar banding patterns and to distinguish between groups of different patterns, for example those from different sources. FPCA thereby finds typical shapes of these functions and forms groups based on the similarity of the empirical functions to the typical functions, based on the difference of the individual functions from these typical functions.

2.2. Approach

The analysis approach described here is based on several DGGE profiles of 16S rRNA gene fragments amplified from replicate environmental samples from different locations or treatments, although the method applies equally to T-RFLP, other genes or other fingerprinting data. For each replicate, the data resulting from these gels consist of a vector of measurements reflecting the intensity of the band from one gel and a second vector of the locations of all bands along the gel. The aim is to identify groups of replicates with similar banding patterns as expressed by band intensity relative to the location. Data analysis consists of three main steps. An initial step yields representation of the banding patterns as a continuous function of intensity. This requires the application of smoothing techniques which are usually provided by the software. In the current approach we used a b-spline smoothing approach as described in Ramsay and Silverman (2005). Multivariate methods for functional data approaches, in particular FPCA, are then applied to the resulting functions. This yields a small number (in the current application usually two) of typical shapes of functions (the so-called functional principal components) that best reflect the variation in shapes among the empirical functions. These typical functions indicate, for example, areas along the gel in which there is the strongest variation in expression among the samples, i.e. areas where some samples are highly expressed and others are not. In the final step, the similarity of the functions derived for each of the samples to the typical functions is assessed, based on the scores of the individual samples on the functional principal components. Hierarchical cluster analysis using Ward's method (Everitt et al., 2001) is applied to these scores to identify groups of similar functions, i.e. similar banding patterns. The analysis was run using the free S-Plus library for functional data analysis (Clarkson et al., 2005). A similar add-on package "fcd" exists for the free software package "R" and can also be used for the analysis.

2.3. Simulation study

In order to assess the suitability of the proposed approach, a detailed simulation study was run to compare DGGE profiles of samples from two artificially generated communities. Each profile consisted of 100 bands, with 20 replicates of each sample; they were simulated to form two groups of 20 samples each. To facilitate comparison, both communities had the same species abundance curve, generated from a log-normal distribution. Any method can easily distinguish communities with strongly differing abundance structures. In this simulation study we were interested in analysing only the impact of those aspects that are likely to cause problems, i.e. noise, low spatial clearness, the number of bands in common and the amount of "smear". Therefore abundance was held constant and only the strength of the factors of interest was varied.

The relative abundances (band intensities) were arranged randomly along a (virtual) gel to generate two different community structures. These were then subjected to different types of artificially generated noise. Random noise (generated from a normal distribution with zero mean) was added to the relative abundances of replicates for each of the groups in the gels. By varying the standard deviation of the normal distribution from which the random noise was generated, the strength of the noise was varied to mimic different degrees of within group variation. The strength of the noise was classified as "weak" if the normal distribution had a standard deviation of 1, as "medium" with a standard deviation of 5 and as "strong" with a standard deviation of 10.

In addition, the degree of spatial clearness was varied by changing the degree of "spatial similarity" between the two groups, i.e. by preferentially locating the most intense bands in similar or dissimilar areas along the gel. This addressed comparisons between communities with sequences with different migration characteristics. The spatial clearness was classified as "low" if the probability that an intense band was preferentially located in a dissimilar area was 0.3%, as "medium" if the probability was 0.6 and as "strong" if the probability was 1.The number of bands in common was varied by fixing the location of a certain fraction of the most intense bands between the two communities. In the simulations the two groups had either 25%, 50% or 75% of the bands in common. Furthermore, the probability of a neighbouring band being (erroneously) detected rather than, or in addition to the "true" band (referred to as "smear"), was varied in different simulation runs. This was achieved by randomly switching the band labelling with varying probability.

The simulation study was run within the commercial statistical software package Splus, which has a specific function (rnorm) to generate random samples from a normal distribution with a given mean and standard deviation.

As the study was dealing with artificial data, the true allocation of each of the 2×20 replicates to either of the groups was known, enabling assessment of the quality of the results, by counting the number of misclassifications for each run. For each combination of strength of noise, spatial clearness, number of bands in common and degree of smear, the average number of misclassifications over 100 runs was calculated for the suggested FPCA approach and then compared to those for the standard approach, i.e. (non-functional) PCA. The standard PCA was also run in the commercial software package S-plus, using the princomp function. To do this, the 100 bands in each of the 20 samples from the 2 groups were treated as if they were independent measurements on 100 variables and the resulting 40×100 data matrix was analysed with the function princomp.

2.4. Application

To test the FPCA method with sample, rather than virtual data, DGGE profiles were analysed from soil samples obtained from cores (5 cm diameter) taken from the uppermost 2–5 cm of an experimental site, '18 Acres field', in Berkshire, UK (grid reference SU 889 737) (Baker et al., in press). Soil was homogenised by sieving (3.25 mm mesh size) and stored at −20 °C until use. Nucleic acids were extracted from 0.5 g soil samples (Griffiths et al., 2000) and bacterial 16S rRNA genes were amplified using a nested PCR approach, with primary amplification employing primers 27f and 1492r (Lane, 1991) and secondary amplification with primers P3 and P2 (Muyzer et al., 1993) as described by Rangel-Castro et al. (2005). PCR products were analysed using DGGE as described previously (Nicol et al., 2005) but with a denaturing gradient of 30–75% denaturant and run in 6.5 l TAE buffer for 999 min at 90 V. Triplicate soil samples were analysed. Samples 1–3 were from individual soil cores, samples 4–6 were from 5 cores pooled together, samples 7–9 from 25 cores pooled together and samples 10–12 were from a further 5 pooled cores. The resulting DGGE consisted of 12 lanes containing 95 bands in total. Band intensities and positions were determined and recorded using PHORETIX 1D gel analysis software (v4,0, Phoretix International, Newcastle upon Tyne, UK) and used as the data matrix for the multivariate analysis. The data were given a functional representation and were smoothed by penalising roughness based on b-splines as described above.

Table 1
Average number of misclassifications for functional PCA approach (left hand figure) and standard PCA approach (right hand figure).

Within group variation	Spatial clearness					
	High		Medium		Low	
	FPCA	PCA	FPCA	PCA	FPCA	PCA
Small number of bands in common (25%)						
Weak	0.28	2.03***	0.45	3.06***	0.95	3.01***
Medium	0.33	2.65***	1.74	2.31***	3.28	5.59**
Strong	0.67	2.59***	1.8	4.68***	3.4	7.1***
Medium number of bands in common (50%)						
Weak	0.65	7.51***	1.12	7.37***	0.8	8.5***
Medium	1.03	5.93***	2.28	8.49***	2.5	9.29***
Strong	1.28	8.3***	2.27	8.8***	4.24	11.05***
Large number of bands in common (75%)						
Weak	0.57	3.11***	2.17	4.79***	1.66	7.07***
Medium	0.34	3.84***	2.95	5.49***	3.84	7.58***
Strong	0.45	3.47***	3.56	6.12***	4.4	8.77***

** $p<0.01$, *** $p<0.001$ with 25% probability of smear.

3. Results

3.1. Simulation study

The current simulation study used 20 replicates for each of the two groups considered, as detailed in the previous section. Sensitivity analysis showed that results similar to those described below were obtained with smaller samples sizes and are not reported separately.

In the absence of smear, both the PCA and FPCA approaches produced almost identical results, with few misclassifications, on average less than 1 misclassification for all simulations. Since the methods cannot be distinguished in this respect, these results are not discussed here in greater detail. However, in the analysis with both low levels of smear (probability of smear 25%, Table 1) and with high levels of smear in the gel (probability of smear 50%, Table 2), the FPCA method was much better than the PCA approach in all cases. The number of misclassifications was significantly smaller in all cases; only in one case did the statistical test yield $p<0.01$ and in all other cases $p<0.001$. Thus, FPCA produced significantly fewer misclassifications (all p-values based on paired two sample t-tests).

The overall mean number of misclassifications over all combinations of settings was 2.07 for FPCA and 7.41 for PCA. A paired two-sample t-test was applied to assess whether the numbers were significantly different. This resulted in a very low p-value ($t=17.53$, $df=53$, p-value $<2.2\times10^{-16}$). That is, overall, significantly fewer misclassifications occur when FPCA was applied than with PCA. With greater smear on the gel, the number of misclassifications was significantly increased for PCA (mean 5.98 versus 8.85, one-sided two-sample t-test: $t=-4.304$, $p<0.001$). However, there was no significant difference between the two levels of smear for FPCA (mean 1.81 versus 2.33, one-sided two-sample t-test: $t=-1.28$, $p>0.05$). Hence, smear had only a negligible effect on the quality of the results for FPCA but strongly affected the PCA results. In the worst case scenario (strong smear, high noise and many bands in common), the FPCA approach produced 6.8 (34%) misclassifications, while PCA generated 12.1 (62.5%). In addition, almost all results for PCA with high smear (and many for weak smear) were worse than the result for this worst case scenario for FPCA.

Table 2
Average number of misclassifications for functional PCA approach (left hand figure) and standard PCA approach (right hand figure).

Within group variation	Spatial clearness					
	High		Medium		Low	
	FPCA	PCA	FPCA	PCA	FPCA	PCA
Small number of bands in common (25%)						
Weak	0.38	4.53***	0.9	6.07***	0.94	8.11***
Medium	0.64	4.79***	2.71	6.2***	3.39	8.61***
Strong	0.96	5.69***	2.32	7.1***	4.19	8.44***
Medium number of bands in common (50%)						
Weak	0.73	9.24***	1.85	10.57***	1.77	11.53***
Medium	0.92	10.2***	2.15	10.7***	2.45	10.58***
Strong	1.74	10.02***	3.58	11.09***	4.04	12.67***
Large number of bands in common (75%)						
Weak	0.6	6.45***	3.21	9.72***	2.82	11.37***
Medium	0.39	6.16***	3.62	9.26***	4.45	11.78***
Strong	0.75	6.49***	4.6	10.12***	6.8	12.05***

** $p<0.01$, *** $p<0.001$ with 50% probability of smear.

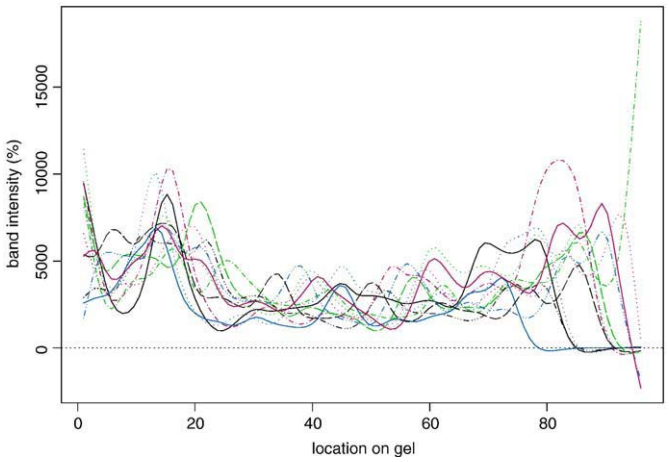

Fig. 2. Smoothed functional representation of the 12 samples derived from a study of DGGE profiles from an experimental site, '18 Acres field', in Berkshire, UK. For further details, see Baker et al. (in press).

3.2. Application

FPCA was applied to the smoothed function, illustrated in Fig. 2. The first three principal components explained 75.1% of the variation in the data (first PC: 34.5%, second PC: 23.9%, third PC: 16.7%). Further PCs were not considered, since the additional amount of variation explained from the fourth PC onwards was rather low (<8%). The plot of the function PC (Fig. 3) shows that the functions varied most towards the end of the gel for bands around (second PC) and above

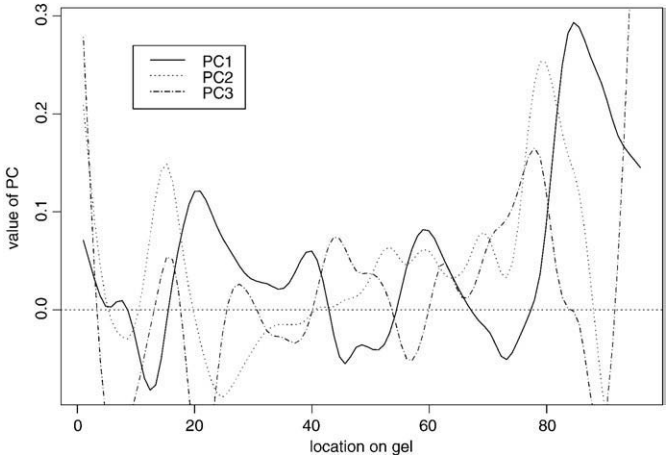

Fig. 3. Plot of the first three functional principal components (PCs) calculated for the 12 samples derived from a study of DGGE profiles from an experimental site, '18 Acres field', in Berkshire, UK (see Baker et al. in press).

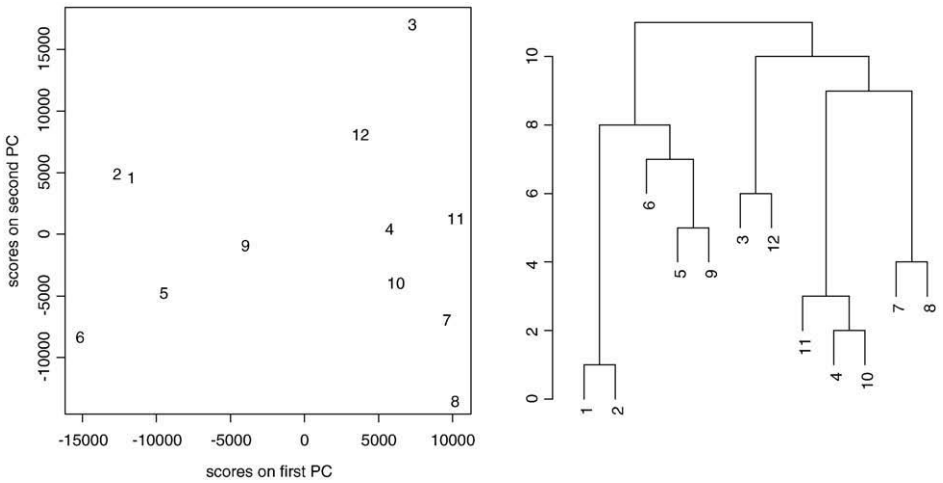

Fig. 4. Plot of the scores of the samples on the first two principal components (left hand side) and dendrogram resulting from a cluster analysis (Ward's algorithm) of the scores on the first two principal components after a functional principal component analysis for the 12 samples derived from an experimental site, '18 Acres field', in Berkshire, UK (see Baker et al. in press).

band 80 (first PC), as marked on the *x*-axis. Note that the values of the functions do not have a direct interpretation and hence can be positive or negative as they do not reflect band intensity. Strong variation was also observed around bands 20 and 40.

A plot of the scores of the samples on the first and second PC (see Fig. 4, left hand side) shows that samples along the first PC axis are split into two clear groups, one including samples scoring high on the first PC and another scoring low (negative) on the first PC. The second PC does not produce results that are as clear-cut as the first PC, which is expected, as it explains less variation. This is reflected in the dendrogram in Fig. 4 (right hand side) which clearly shows two distinct groups of samples, those scoring high on the first PC (group 1 = 3, 4, 7, 8, 10, 11 and 12) and those scoring low on the first PC (group 2 = 1, 2, 5, 6, 9). Samples in group 1 tend to have quite high relative intensity around bands 20, 40, 60 and 82, whereas those in group 2 have rather low expressions in those bands.

A standard PCA was also run for these data and the results are summarised in Fig. 5. The first two PCs explain approximately 31% of the variation; consideration of the first seven principal components would be required to explain at least 75.1% of the variation (as was achieved by the FPCA). Apparently, the PCA struggles to find a clear structure in the data which is clearly picked up by the FPCA. It is very obvious that the results of the PCA analysis differ very strongly from those of the FPCA. For example, while samples 10 and 11 seem to be rather similar in both analyses, sample 4 appears to be very different to them in the PCA analysis but very similar in the FPCA analysis. The dendrogram structure also varies strongly and the grouping structure is much less clear, a further indication that the data structure has not been clearly identified by the PCA. When the scores of the first seven principal components are included in the analysis (as suggested by the percentage of variation explained by the principal components) the dendrogram changes but still differs from the FPCA results (not shown). It is not possible to prove which of the two analyses presents the true structure for this real data set, in the absence of additional information. However, taking into account the conclusions from the simulation study the results from the FPCA are likely to be more reliable.

4. Discussion

A robust analysis of the complexity of microbial communities in environmental samples is a fundamental step towards an increased

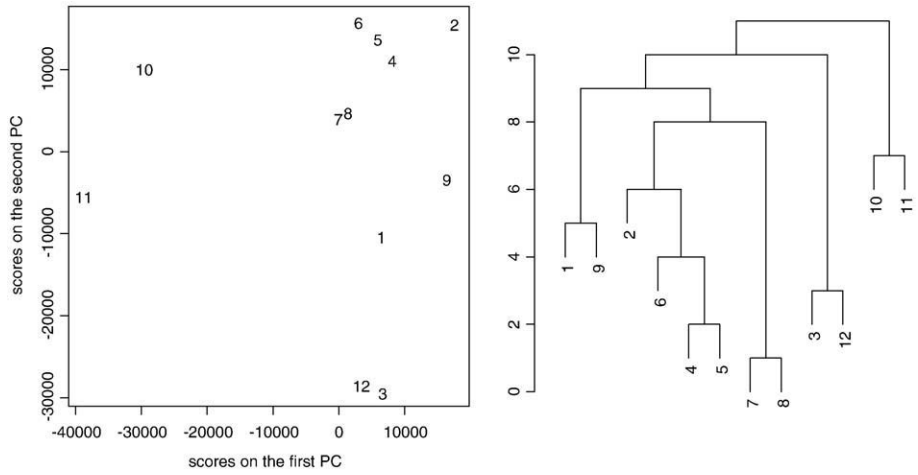

Fig. 5. Plot of the scores of the samples on the first two principal components (left hand side) and dendrogram resulting from a cluster analysis (Ward's algorithm) of the scores on the first two principal components after a classical principal component analysis for the 12 samples derived from an experimental site, '18 Acres field', in Berkshire, UK (see Baker et al. in press).

understanding of how communities are affected by physicochemical processes and climatic changes in different ecosystems (Blackwood et al., 2003; Ramette, 2009) at different levels of resolution, dependent on which genes or molecular biomarkers are used for a particular fingerprinting technique (Nakatsu, 2007). The number of ordination methods used to make such analyses has increased in the past few years (Ramette, 2007). However, most of these methods fail to separate samples that are similar but which present subtle differences that may provide important information on the mechanisms controlling microbial diversity in a particular ecosystem. Furthermore, inherent technical problems in fingerprinting methods, such as the presence of smear in DGGE, can cause further difficulties for community analysis. In most cases, considerable data handling is required to remove this background from the real data and several strategies have been adopted to improve data processing and analysis of profiles using T-RFLP to compare microbial communities from different environments (Blackwood et al., 2003; Fierer and Jackson, 2006; Culman et al., 2008). This paper proposes and assesses a new approach, functional PCA, for analysis of data derived from molecular fingerprinting techniques, particularly DGGE, and compares it to a classical multivariate analysis method, principal component analysis, in a simulation study where similarities and differences among artificially generated samples are known. Analysis shows that both the standard PCA and the functional PCA approaches produced similarly reliable and robust results under 'ideal' but unrealistic conditions, even in the presence of noise. Under conditions that are more typical of real situations, where false neighbouring bands are taken into account in the analysis, major problems were experienced with PCA. In particular, the method misclassified a significantly higher number of samples than the suggested functional approach. These issues arise mainly because PCA treats each band as being independent of neighbouring bands and it therefore does not take into account the structure and location of the bands along the gel. As a result, information on neighbourhood structures is not included in the analysis. FPCA, however, makes this distinction, as the observations derived from the banding patterns are treated as functions of the location along the gel. The shape of these functions is classified such that samples with common false neighbouring bands are considered more similar than samples showing false bands at a random location elsewhere. Consequently, the number of misclassifications is quite low. Furthermore, similar samples for which the main source of variation is band intensity will be better separated by FPCA than PCA, as the final mathematical function of the profile will be modified to a greater extent than single, discrete, points on a PCA.

In contrast to the simulation study, the level of background or smear in real fingerprinting data, such as that obtained from amplified 16S rRNA gene fragments from soil samples and analysed by DGGE, is not known. In such analyses, it is often difficult to guarantee that the data in the final matrix have been collected without artefacts. It is hence advisable to apply FPCA methods to analyse real data as the method is less likely to have misclassified the samples as shown in this paper. A potential disadvantage of the FPCA approach is that it might be incapable of detecting differences between groups of samples that truly differ only subtly by neighbouring bands. However, through choosing a small degree of smoothing these can still be distinguished, but in many applications this might be easily masked by technical noise resulting from smear, and results should be interpreted carefully. The application of both methods to a real data set also showed distinctly different results for the two analysis methods, emphasising further the relevance of this study to applications.

In conclusion, FPCA provides a better approach to microbial community analysis based on fingerprinting techniques, such as DGGE and T-RFLP, and has the potential to significantly reduce problems arising from smearing. The approach can be used to assess the significance of communities at different locations and subjected to different environmental conditions and more general functional data analysis methods, such as functional linear models, could be used to relate functional and non-functional data to each other. In these models the functional data can be either an explanatory variable (e.g. in attempting to relate the structure of the microbiological community to a measure of soil quality) or an outcome variable (e.g. where soil characteristics are sought to explain the structure of the microbiological community).

Acknowledgements

JIR-C acknowledges funding from UK Population Biology Network (UKPopNet), funded by NERC, and KLB acknowledges receipt of a BBSRC CASE studentship with Syngenta.

References

Baker, K.L., Langenheder, S., Nicol, G.W., Ricketts, D., Killham, K., Campbell, C.D. and Prosser, J.I., in press. Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. Soil Biology and Biochemistry. doi:10.1016/j.soilbio.2009.08.010.

Blackwood, C.B., Marsh, T., Kim, S.-H., Paul, E.A., 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. Applied and Environmental Microbiology 69 (2), 926–932.

Cardenas, E., Tiedje, J.M., 2008. New tools for discovering and characterizing microbial diversity. Current Opinion in Biotechnology 19 (6), 544–549.

Clarkson, D.B., Fraley, C., Gu, C.C., Ramsay, J.O., 2005. S+Functional Data Analysis, User's Manual for Windows ®. Springer.

Culman, S.W., Gauch, H.G., Blackwood, C.B., Thies, J.E., 2008. Analysis of T-RFLP data using analysis of variance and ordination methods: a comparative study. Journal of Microbiological Methods. 75 (1), 55–63.

Everitt, B., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold, London.

Fierer, N., Jackson, R.B., 2006. The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences. 103 (3), 626–631.

Fromin, N., Hamelin, J., Tarnawski, S., Roesti, D., Jourdain-Miserez, K., Forestier, N., Teyssier-Cuvelle, S., Gillet, F., Aragno, M., Rossi, P., 2002. Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns. Environmental Microbiology. 4 (11), 634–643.

Griffiths, R.I., Whiteley, A.S., O'Donnell, A.G., Bailey, M.J., 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Applied and Environmental Microbiology 66 (12), 5488–5491.

Illian, J.B., Benson, E., Crawford, J., Staines, H.J., 2006. Principal component analysis for spatial point patterns. In: Baddeley, A., Gregori, P., Mateu, P., Stoica, R., Stoyan, D. (Eds.), Case studies in spatial point process modelling. Springer, New York.

Lane, D.J., 1991. 16S/23S rRNA sequencing. In: Stackebrandt, E., Goodfellow, M. (Eds.), Nucleic acid techniques in bacterial systematics. John Wiley and Sons, New York, pp. 115–175.

Liu, W.-T., Marsh, T.L., Cheng, H., Forney, L.J., 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Applied and Environmental Microbiology 63, 4516–4522.

Muyzer, G., 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. Current Opinion in Microbiology 2, 317–322.

Muyzer, G., Smalla, K., 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology 73 (1), 127–141.

Muyzer, G., De Waal, E.C., Uitterlinden, A.G., 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59:3, 695–700.

Nakatsu, C.H., 2007. Soil microbial community analysis using denaturing gradient gel electrophoresis. Soil Science Society of America Journal 71 (2), 562–571.

Nicol, G.W., Tscherko, D., Embley, T.M., Prosser, J.I., 2005. Primary succession of soil Crenarchaeota across a receding glacier foreland. Environmental Microbiology 7:3, 337–347.

Prosser, J.I., Embley, T.M., 2002. Cultivation-based and molecular approaches to characterisation of terrestrial and aquatic nitrifiers. Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology 81, 165–179.

Ramette, A., 2007. Multivariate analyses in microbial ecology. FEMS Microbiology Ecology 62 (2), 142–160.

Ramette, A., 2009. Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. Applied and Environmental Microbiology. 75 (8), 2495–2505.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, 2nd edition. Springer, Berlin.

Rangel-Castro, J.I., Killham, K., Ostle, N., Nicol, G.W., Anderson, C., Scrimgeour, C.M., Ineson, P., Meharg, A., Prosser, J.I., 2005. Stable isotope probing analysis of the influence of liming on root exudate utilisation by soil microorganisms. Environmental Microbiology 7, 828–838.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Smalla, K., Oros-Sichler, M., Milling, A., Heuer, H., Baumgarte, S., Becker, R., Neuber, G., Kropf, S., Ulrich, A., Tebbe, C.C., 2007. Bacterial diversity of soils assessed by DGGE, T-RFLP and SSCP fingerprints of PCR-amplified 16S rRNA gene fragments: Do the different methods provide similar results? Journal of Microbiological Methods 69 (3), 470–479.

Thies, J.E., 2007. Soil microbial community analysis using terminal restriction fragment length polymorphisms. Soil Science Society of America Journal. 71 (2), 579–591.

Torsvik, V., Øvreås, L., 2002. Microbial diversity and function in soil: from genes to ecosystems. Current Opinion in Microbiology 5 (3), 240–245.

Sunnucks, P., Wilson, A.C., Beheregaray, L.B., Zenger, K., French, J., Taylor, A.C., 2000. SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Molecular Ecology 9 (11), 1699–1710.