Available online at www.sciencedirect.com
ScienceDirect

International Journal of Food Microbiology 120 (2007) 95 – 99

INTERNATIONAL JOURNAL OF Food Microbiology

www.elsevier.com/locate/ijfoodmicro

Use of multivariate statistics for 16S rRNA gene analysis of microbial communities

K. Rudi [a,b,*], M. Zimonja [a], P. Trosvik [a], T. Næs [a]

[a] MATFORSK Norwegian Food Research Institute, Osloveien 1, NO-1430 Ås, Norway
[b] Hedmark University College, Holsethgt. 22, NO-2318 Hamar, Norway

Abstract

Understanding dynamic processes and diversity in microbial communities is of key importance for combating pathogens and for stimulating beneficial bacteria. We have addressed these challenges utilising multivariate statistics for analyses of microbial community structures. We based our microbial community analyses on 16S rRNA gene data. This gene is by far the most widely applied genetic marker for phylogenetic and microbial community studies. Both probe and clone library data were analysed. We analysed the clone library data using a newly developed coordinate-based phylogenetic approach. By using coordinates, we avoid both DNA sequence alignments and the need for definition of operational taxonomic units (OTUs). The basic principle is to transform the sequence data to frequencies of multimers (short sequences of $n=2$ to 6), and then to use principal component analyses (PCA) for data compression into an orthogonal coordinate space. We used our coordinate method for global 16S rRNA gene analyses of prokaryotes. When comparing microbial communities, it is often important to determine the relationship between the microflora and knowledge about the samples analysed. We used partial least square regression (PLSR) to relate physical/chemical properties to microbial community composition. This was done by analysing both probe and clone library data using the effect of modified atmosphere packaging (MAP) on fish microflora as an example. We are currently investigating approaches to describe dynamic microbial community interactions. Our ultimate goal is to understand and model the main dynamic interactions in complete microbial communities.

Keywords: Microbial community; Multivariate statistics; 16S rRNA gene

1. Introduction

It has become apparent in recent years that both the beneficial and harmful effects of bacteria are often due to bacterial interactions in complex communities (Bell et al., 2005). Bacterial monocultures hardly ever exist in nature, except for perhaps in some extreme environments. The main reason for investigating bacteria in monocultures is the lack of tools for analyses in complex environments. The advancement of novel technologies to generate and analyse data from microbial communities, however, is transforming our view of microbial interactions (Cowan et al., 2005).

Of particular importance is the recent adaptation of multivariate statistical tools from chemomentrics to microbial community analysis. The use of multivariate statistics has now been demonstrated with respect to environment (Mouser et al., 2005), food (Blaiotta et al., 2004; Pepe et al., 2004; Rudi et al., 2004) and humans (Wang et al., 2004). Multivariate statistical analyses of microbial communities, however, are still in the infancy.

Currently, a major focus is on describing biodiversity in microbial communities. This research was boosted with the advancement of metagenome sequencing approaches (Venter et al., 2004). The metagenome data, however, are only snapshots of dynamic and spatially separated processes. It is not possible without the dynamic and/or spatial knowledge to determine the relevance or representativity of the samples analysed. Obtaining such knowledge would require approaches that enable microbial community analyses of a large number of samples, and not in-depth analyses of a few samples (Paerl and Steppe, 2003).

The aim of our work is to develop and implement novel 16S rRNA gene tools for microbial community composition and diversity analyses. Our focus is on accurate discrimination and classification of bacteria, and on understanding and describing temporal/spatial differences between communities. In this review we will present a novel alignment independent approach

* Corresponding author. MATFORSK Norwegian Food Research Institute, Osloveien 1, NO-1430 Ås, Norway. Tel.: +47 64 97 02 66; fax: +47 64 97 03 33.
E-mail addresses: knut.rudi@matforsk.no, knut.rudi@hihm.no (K. Rudi).

0168-1605/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijfoodmicro.2007.06.004

for phylogenetic analyses and microbial community comparison. We also introduce multivariate regression to analyse microbial community structures. Finally, we present the potential application of statistical tools for analysing dynamic processes in microbial communities.

2. The genes encoding rRNA

The genes encoding ribosomal RNAs are ancient. They are functionally constant, universally distributed and comprise highly conserved sequence domains interspersed with more variable regions (Woese, 1987). The conserved regions are important for classification of higher taxa, while the variable regions can be used for differentiation between closely related species.

In prokaryotes there are three ribosomal RNA molecules, which have the sizes 5S, 16S and 23S. Initially, the analysis of the diversity of natural microbial populations relied on direct extraction, purification, and sequencing of 5S rRNA molecules from environmental samples, but the information from the 120 nucleotide-long 5S rRNA is relatively limited. An average bacterial 16S rRNA molecule has a length of approximately 1 500 nucleotides, the bacterial 23S rRNA has 2 900 nucleotides, and thus the two contain considerably more information than 5S rRNA. 16S RNA is more experimentally manageable than 23S rRNA, and it has been used extensively to develop the phylogeny of both prokaryotes and eukaryotes. Even though 23S rRNA is larger, the currently available complete sequences are rather poor compared to the coding sequence for 16S rRNA (Ludwig and Schleifer, 1994).

3. General outline of 16S rRNA gene analyses

The first step in 16S rRNA gene microbial community analyses is to purify DNA from all bacteria present in a sample without introducing bias due to e.g. differential lysis or recovery. We used mechanical lysis in combination with magnetic bead-based DNA purification in order to purify DNA from microbial communities (Skanseng et al., 2006). In this way we have obtained both a robust and high throughput DNA purification approach. All bacterial 16S rRNA genes in the samples are subsequently amplified using primers targeting generally conserved regions in the 16S rRNA gene. The ratio between 16S rRNA gene copies for different bacteria is (in theory) conserved during the amplification due to the use of a universally conserved primer pair (Rudi, 2003).

The final step in the analysis is to detect and quantify 16S rRNA genes from the different bacteria present in the mixed amplification product. This can be done using probe based methods (Rudi et al., 2002), cloning and DNA sequencing (Rudi et al., 2004), or electrophoresis-based methods such as T-RFLP (Wang et al., 2004) and denaturing/temperature gradient gel electrophoresis (DGGE/TGGE)(Ercolini, 2004; Muyzer and Smalla, 1998). The detection methods have different fields of application. Probe-based methods are generally used if one knows which bacteria to search for, while cloning and DNA sequencing is used if one needs specific information from unknown samples. DGGE or T-RFLP, on the other hand, is used for screening purposes and pattern recognition (Theron and Cloete, 2000). A schematic outline of the process is shown in Fig. 1.

4. Multivariate statistical tools

There is a wide range of multivariate statistical tools available with potential application for microbial community analyses. A more extensive description of the tools applied here can be found in Martens and Næs (1989). Our work is mainly based on the use of principal component analysis (PCA) for data compression and partial least square regression (PLSR) for multivariate regression analyses.

PCA is a projection method that helps visualise the main information contained in large data tables (Fig. 2). PCA

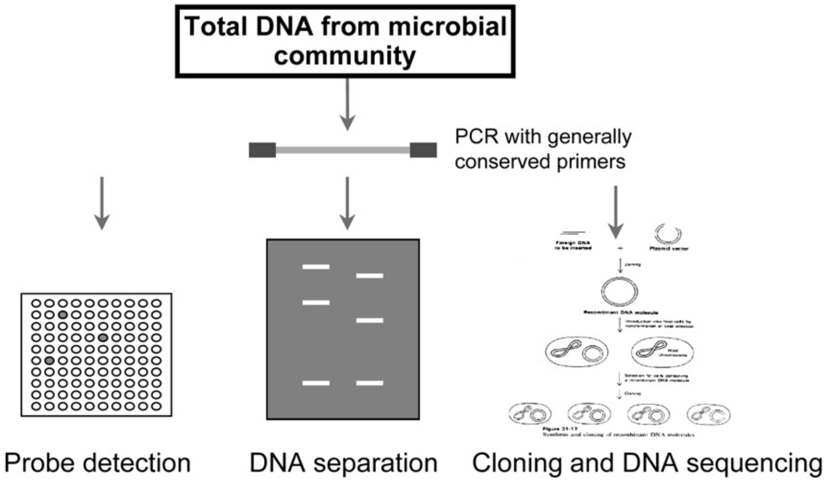

Total DNA from microbial community
PCR with generally conserved primers
Foreign DNA to be inserted
+
Plasmid vector
Recombinant DNA molecule
Figure 31-17
Synthesis and cloning of recombinant DNA molecules
Probe detection
DNA separation
Cloning and DNA sequencing

Fig. 1. Schematic outline of microbial community analysis process. The first step in microbial community analyses is DNA purification. This step is crucial because biases here will affect all subsequent analysis. The next step is to PCR amplify the 16S rRNA gene from all bacteria present in the sample using universally conserved primers. There are several alternative detection methods depending on the problem investigated. Probe-based detection methods are used if one knows the important bacteria. DNA separation-based methods are used if the aim is to determine bacteria profiles, while cloning and DNA sequencing are used if in-depth analyses of bacterial communities are needed.

involves a mathematical procedure that transforms a number of (possibly) correlated variables (in our case: microbial community data) into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The result is a new set of variables that represent linear combinations of the original variables that are uncorrelated and reflect the most important structure of the data. Values for each sample projected onto these "loadings" are then calculated and called "scores".

PLSR is a bilinear modelling method in which information in the original *X* data (in our case microbial community composition) is projected onto a small number of underlying ("latent") variables called PLS components (Fig. 3). The *Y*-data (in our case physical/chemical properties of the samples) are actively used in estimating the "latent" variables to ensure that the first components are those that are most relevant for predicting the *Y*-variables (Bjornstad et al., 2004).

5. Alignment independent 16S rRNA gene analyses

Alignment-based approaches are not well suited for analyses of microbial community clone frequency data. The main reason is that the computer operation time for aligning DNA sequences by dynamic programming increases exponentially with the number of taxa analysed (exhaustive approaches). Furthermore, the statistical testing for microbial community comparison is also very difficult (Curtis and Sloan, 2004).

We have recently developed an alignment independent phylogenetic approach which is based on PCA (Rudi et al., 2006). The first step in our method (denoted AIBIMM; alignment independent bi-linear multivariate modelling) is to transform DNA sequence data into DNA *n*-mer frequencies. The *n*-mer frequency data are obtained by sliding a window of size *n* (i.e. *n* nucleotides) and are identified by the base combination present for each step. The number of contributions for each of the 4^n combinations is then counted. The second step is based on using these data as input for PCA (Fig. 4).

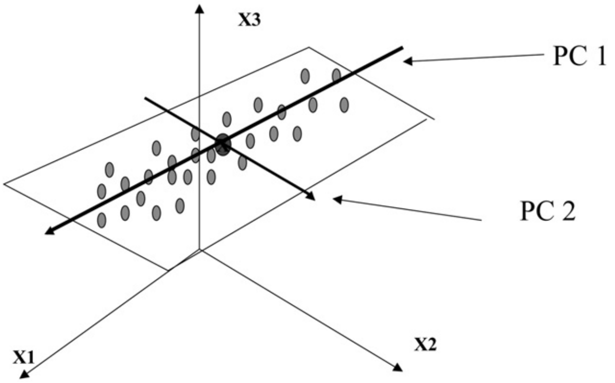


Fig. 2. Illustration of the PCA principle. A set of objects (dots) is described with the three variables *X*1, *X*2 and *X*3. The first principal component (PC1) is the line explaining most of the variance. The second principal component (PC2) is the line orthogonal to PC1 that explains the most of the residual variance.

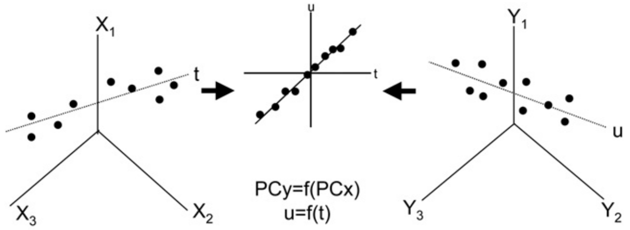


Fig. 3. Illustration of the PLSR principle. PLSR is used to find the underlying correlation between two data tables *X* and *Y*, where *Y* is described as a function of *X*. PLSR utilises both tables of data in identifying this correlation.

We have developed the computer programme PhyloMode for conducting AIBIMM analyses. The program can be downloaded free of charge from www.matforsk.no/web/sampro.nsf/downloadE/Microbial_community. The PhyloMode programme contains two basic modules. The first module transforms DNA sequences into multimer data ($n=1$ to $n=6$). The input is a file in FASTA format. The output from the module can be exported in tab delimited text for advanced multivariate statistical analyses by software packages such as Unscrambler (Camo Inc., Corvallis, Oregon). The PhyloMode software also includes a module for principal component analyses and 2D visualisation of both the score and loading plots. Finally, the program has an option for creating dendrograms based on the principal component data using single, centroide or complete linkage. The linkage data are exported in a format compatible with the freeware TreeView (taxonomy.zoology.gla.ac.uk/rod/treeview.html), which is a software package for drawing phylogenetic trees.

There are more than 200,000 environmental 16S rRNA gene sequences, and the number is estimated to double every 7 months (Frank and Pace, 2005). Frequency analysis of groups of organisms that constitute operational taxonomic units (OTUs) in clone libraries is currently the most widely used approach for studying structures in microbial communities (Rudi et al., 2006). The problem, however, is that there are no natural species barriers for asexual prokaryote species, and thereby no rationale criteria for OTU definitions (Gevers et al., 2005). Furthermore, most microbial species are actually not yet characterised, making it very difficult to define OTUs.

Our concept for microbial community comparisons is that we base the analyses on describing the evolutionary relatedness between bacteria using AIBIMM. Then, we use densities within the AIBIMM coordinate space to compare bacteria from different communities. In this way we avoid having to define OTUs. The comparisons are also very efficient with respect to computer operation time (CPU) since it only involves a direct comparison of taxon densities.

6. Use of multivariate regressions

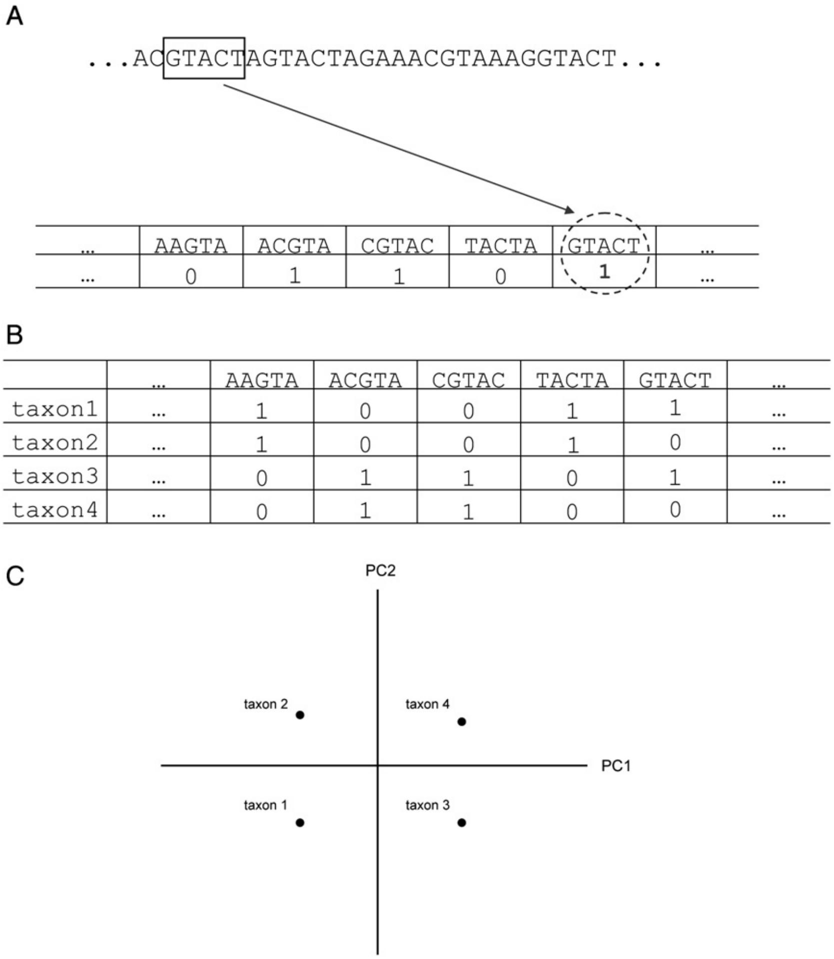
In microbial community analyses, multivariate regression techniques can be used to relate physical/chemical conditions to microbial community structures. The main application is to identify complex correlations in the data involving multiple bacteria. Microbial communities are often composed of several hundred species making it very difficult to identify correlations by traditional means.

There are numerous ways to generate information from microbial communities (Theron and Cloete, 2000). We have focused on using either clone frequency data or probe signal intensities. Probe signals can be generated using microarrays, capillary gel electrophoresis, quencher extension or other techniques. Both the clone frequency and the probe data can be treated in the same way with respect to microbial community analyses. That is, both approaches describe the relative abundance of different groups of bacteria in a microbial community. Groups of bacteria are treated as *X* variables, while physical or chemical properties are treated as *Y* variables in microbial community comparisons using multivariate regression. Multivariate regression can then be used to determine the variation in microbial communities that can be explained by the physical/chemical conditions.

We applied the multivariate regression technique PLSR to investigate the correlation between fish matrix (salmon and coalfish) and storage time in a modified atmosphere (MAP) at low temperatures (1 and 5 °C) (Rudi et al., 2004). Our analyses showed that there were very large differences in the microflora for salmon and coalfish. The most prominent features were the association of coalfish with *Photobacterium* and salmon with *Brochothrix* and *Carnobacterium*. There was also a clear association of storage time with *Photobacterium* for coalfish, while for salmon there was positive association for *Carnobacterium* and negative association for *Brochothrix* with storage time. Thus, using PLSR or other multivariate regression methods we now have the possibility to better understand the effect of physical/chemical conditions on the microflora.

7. Multivariate analyses of microbial community time series data

Microbial communities are highly dynamic systems, and understanding these dynamics is of crucial importance. Identifying dynamic networks of bacterial interaction is important both for combating pathogens, and for the use of probiotic bacteria. Traditional analytical approaches are not suitable to detect these interactions. A long standing problem for the food industry has been to document the effects of probiotic bacteria (Saxelin et al., 2005). Taking into account the dynamic properties of microbial communities, however, could



...	AAGTA	ACGTA	CGTAC	TACTA	GTACT	...
...	0	1	1	0	1	...

	...	AAGTA	ACGTA	CGTAC	TACTA	GTACT	...
taxon1	...	1	0	0	1	1	...
taxon2	...	1	0	0	1	0	...
taxon3	...	0	1	1	0	1	...
taxon4	...	0	1	1	0	0	...

Fig. 4. Alignment independent 16S rRNA gene analysis. (A) The first step in the analysis is to transform DNA sequences into multimer frequencies for each individual taxon. (B) A table of multimer frequencies is created. (C) Finally, redundant multimer frequency information is compressed into PCs.

help identify new mechanisms or strengthen already established hypotheses.

As for all ecological systems, the dynamics of bacterial communities are governed by interaction among community members as well as by the environmental conditions. Given that the generation time for bacteria can be less than half an hour, we would expect microbial ecosystems to be highly dynamic. What is generally investigated, however, are only snap shots of dynamic processes. The main obstacle when it comes to identifying and exploring such intrinsic and extrinsic features is the availability of good time-series data in a scale that is relevant to the organisms under investigation.

Since the current tools to investigate microbial communities have been focused on describing diversity, these techniques are not suitable for analysing multiple samples, which is required in time-series analyses. Thus, there is presently a lack of time-series data that are relevant for microorganisms. We are now evaluating the possibility of generating such data. We are also exploring the possibility of statistical/mathematical tools to analyse time-series data by using generalised additive models (GAM). GAM is a generalisation of multivariate regression models such as PLSR, but instead of parameters, functions are used in the regression. GAM is used for data with complex interactions (for details see (Hastie and Tibshirani, 1990)).

Acknowledgements

We would like to thank the Norwegian Trust for Certain Agricultural Products, Helse Øst and Hedmark Sparebank for the financial support.

References

Bell, T., Newman, J.A., Silverman, B.W., Turner, S.L., Lilley, A.K., 2005. The contribution of species richness and composition to bacterial services. Nature 436, 1157–1160.

Bjornstad, A., Westad, F., Martens, H., 2004. Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR). Hereditas 141, 149–165.

Blaiotta, G., Pennacchia, C., Villani, F., Ricciardi, A., Tofalo, R., Parente, E., 2004. Diversity and dynamics of communities of coagulase-negative staphylococci in traditional fermented sausages. J. Appl. Microbiol. 97, 271–284.

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., Wittwer, P., 2005. Metagenomic gene discovery: past, present and future. Trends Biotechnol. 23, 321–329.

Curtis, T.P., Sloan, W.T., 2004. Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. Curr. Opin. Microbiol. 7, 221–226.

Ercolini, D., 2004. PCR-DGGE fingerprinting: novel strategies for detection of microbes in food. J. Microbiol. Methods 56, 297–314.

Frank, D.N., Pace, N.R., 2005. Another ribosomal RNA sequence milestone-and a call for better annotation. ASM News 71, 501–502.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., de Peer, Y.V., Vandamme, P., Thompson, F.L., Swings, J., 2005. Re-evaluating prokaryotic species. Nat. Rev. Microbiol. 3, 733–739.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall.

Ludwig, W., Schleifer, K.H., 1994. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol. Rev. 15, 155–173.

Martens, H., Næs, T., 1989. Multivariate Calibration. John Wiley & Sons, Chichester.

Mouser, P.J., Rizzo, D.M., Roling, W.F., Van Breukelen, B.M., 2005. A multivariate statistical approach to spatial representation of groundwater contamination using hydrochemistry and microbial community profiles. Environ. Sci. Technol. 39, 7551–7559.

Muyzer, G., Smalla, K., 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek 73, 127–141.

Paerl, H.W., Steppe, T.F., 2003. Scaling up: the next challenge in environmental microbiology. Environ. Microbiol. 5, 1025–1038.

Pepe, O., Blaiotta, G., Anastasio, M., Moschetti, G., Ercolini, D., Villani, F., 2004. Technological and molecular diversity of *Lactobacillus plantarum* strains isolated from naturally fermented sourdoughs. Syst. Appl. Microbiol. 27, 443–453.

Rudi, K., 2003. Application of 16S rDNA arrays for analyses of microbial communities. Recent Res. Dev. Bacteriol. 1, 35–44.

Rudi, K., Flateland, S.L., Hanssen, J.F., Bengtsson, G., Nissen, H., 2002. Development and evaluation of a 16S rDNA array approach for describing complex microbial communities in ready-to-eat vegetable salads packed in modified atmosphere. Appl. Environ. Microbiol. 68, 1146–1156.

Rudi, K., Maugesten, T., Hannevik, S.E., Nissen, H., 2004. Explorative multivariate analyses of 16S rRNA gene data from microbial communities in modified-atmosphere-packed salmon and coalfish. Appl. Environ. Microbiol. 70, 5010–5018.

Rudi, K., Zimonja, M., Naes, T., 2006. Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny. Int. J. Syst. Evol. Microbiol. 56, 1565–1575.

Saxelin, M., Tynkkynen, S., Mattila-Sandholm, T., de Vos, W.M., 2005. Probiotic and other functional microbes: from markets to mechanisms. Curr. Opin. Biotechnol. 16, 204–211.

Skanseng, B., Kaldhusdal, M., Rudi, K., 2006. Comparison of chicken gut colonisation by the pathogens *Campylobacter jejuni* and *Clostridium perfringens* by real-time quantitative PCR. Mol. Cell. Probes 20, 269–279.

Theron, J., Cloete, T.E., 2000. Molecular techniques for determining microbial diversity and community structure in natural environments. Crit. Rev. Microbiol. 26, 37–57.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74.

Wang, M., Ahrne, S., Antonsson, M., Molin, G., 2004. T-RFLP combined with principal component analysis and 16S rRNA gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages. J. Microbiol. Methods 59, 53–69.

Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221–271.