Multivariate Analysis of Complex DNA Sequence Electropherograms for High-Throughput Quantitative Analysis of Mixed Microbial Populations

Pål Trosvik, Beate Skånseng, Kjetill S. Jakobsen, Nils C. Stenseth, Tormod Næs, and Knut Rudi

Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Oslo N-0316, Norway; Matforsk AS, Norwegian Food Research Institute, N-1430 Ås, Norway; Department of Mathematics, University of Oslo, Oslo N-0316, Norway; and Hedmark University College, Hamar N-2318, Norway

Received 18 January 2007/Accepted 3 June 2007

High-throughput quantification of genetically coherent units (GCUs) is essential for deciphering population dynamics and species interactions within a community of microbes. Current techniques for microbial community analyses are, however, not suitable for this kind of high-throughput application. Here, we demonstrate the use of multivariate statistical analysis of complex DNA sequence electropherograms for the effective and accurate estimation of relative genotype abundance in cell samples from mixed microbial populations. The procedure is no more labor-intensive than standard automated DNA sequencing and provides a very effective means of quantitative data acquisition from experimental microbial communities. We present results with the Campylobacter jejuni strain-specific marker gene gltA, as well as the 16S rRNA gene, which is a universal marker across bacterial assemblages. The statistical models computed for these genes are applied to genetic data from two different experimental settings, namely, a chicken infection model and a multispecies anaerobic fermentation model, demonstrating collection of time series data from model bacterial communities. The method presented here is, however, applicable to any experimental scenario where the interest is quantification of GCUs in genetically heterogeneous DNA samples.

As of today, the most popular techniques used for exploring microbial communities are geared toward the description and comparison of diversity and membership (7, 9, 26, 30, 35, 38, 39, 41). While these methods are well suited for providing snapshots of microbial community structure, the population dynamic structure of a community is not a thoroughly developed concept in the field of microbial ecology. In order to capture dynamic interactions in a community of organisms, one needs quantitative time series data collected on a scale that is relevant to the life histories of the biological species in question. In the case of microorganisms, this entails high-frequency enumeration of the different community components.

When working with model microbial communities, an experimenter may select genetic markers that distinguish between the different community members at the taxonomic level of interest. A group of organisms with a shared genotype for such a marker may be said to form a genetically coherent unit (GCU). In this report, we present a novel approach to high-throughput GCU quantification from mixed microbial populations. Using direct mixed template sequencing, we describe the use of multivariate statistical analysis of complex DNA sequence electropherograms for the accurate estimation of relative genotype abundance. An outline of the approach is shown in Fig. 1.

The direct sequencing multivariate-modeling technique was applied to two genetic markers for GCU quantification in different experimental bacterial community scenarios. The first marker was the gltA gene (encoding citrate synthase), which was used as a marker for seven strains of pathogenic Campylobacter jejuni. The genetic sequences of gltA are closely related in these seven strains, and polymorphic sites are represented by a few single nucleotide polymorphisms (SNPs). The second marker was the 16S rRNA gene, which is universally conserved, with a high degree of polymorphism across bacterial lineages. We used the 16S rRNA gene as a marker for seven more distantly related GCUs representing common gut bacteria. For both of these genetic markers, we computed statistical models for GCU quantification in either community scenario.

We also present results from the analysis of two sets of direct community sequence data from mixed population experiments. In both cases, this analysis provided us with time series population data, thus demonstrating the usefulness of our approach. For the first set, we used the gltA gene model to monitor the colonization kinetics of different C. jejuni strains in the chicken cecum. Second, the 16S rRNA gene model was used to describe the dynamics in a 9-h anaerobic fermentation with three bacterial species commonly found in the gastrointestinal tracts of humans and animals.

MATERIALS AND METHODS

General outline of the direct sequencing approach. The selection of an appropriate genetic marker is determined by the experimental setting, i.e., the genetic diversity within the system under study (Fig. 1, step 1). The marker must
1. Selection of appropriate genetic marker and modeling approach.

2. DNA sequencing and pre-processing of spectral sequence data.


FIG. 1. Flow chart outlining the general methodology. The selection of a genetic marker and modeling approach is determined by the experimental setting. If PLSR is to be performed, the experimenter must supply a set of training data for model computation. If an MLR approach is preferred, the electropherograms for the selected genetic marker for the GCUs involved in the experiment must be supplied, preferably along with a test set of data for estimation of prediction accuracy and bias. Once the model has been computed and properly validated, it can be applied to real-time mixed population experiments.

be conserved throughout the system yet have enough sequence diversity to distinguish between the GCUs present. DNA sequence training data for the chosen marker are produced (Fig. 1, step 2), and a multivariate model for quantification is computed and validated with these data (Fig. 1, step 3). The multivariate statistical techniques used here are those of the multiple linear regression (MLR) method according to the linear mixture model and partial least-squares regression (PLSR). Once a satisfactory model has been obtained, a mixed population experiment can be conducted (Fig. 1, step 4). Upon bulk DNA extraction from an experimental community sample, the pertinent genetic marker is amplified by PCR using primers targeting conserved regions of the marker in question (37). In the absence of significant amplification efficiency bias, the resulting mixture of homologous DNA fragments should approximate the GCU composition of the community from which the sample was collected (12). The resulting mixture is sequenced directly with an automated sequencer using a standard technique of cyclic fluorescence labeling and chain termination with a conserved primer. The sequencing apparatus will then produce an emission spectrum (electropherogram) characteristic of the mixture in question, carrying in it quantitative information about the original sample composition. This information is extracted using the previously developed regression model.

Generation of mixed sequence spectra for multivariate modeling. PCR amplification of the gltA genes was carried out using the primers gltTF (5′-CGT CCT TTT GAT TCT TTY CCT GAT AAT-3′) and gltTR (5′-GAA GTW GAA GCA TTT TGY TCA TGA TCT G-3′) (3). The 25-μl PCR mixture contained 1× Hot Start buffer (Finnzymes, Espoo, Finland), 200 μM of a deoxynucleoside triphosphate (dNTP) mixture, 1 U of DyNAzyme II Hot Start DNA polymerase (Finnzymes), 0.2 μM of each primer, and 1 μl of DNA. The amplification profile was an initial step of 95°C for 10 min and then 35 cycles of 95°C for 30 seconds, 50°C for 2 min, and 72°C for 30 seconds, and a final extension at 72°C for 7 min. The relatively high number of PCR cycles used was required due to the low colonization numbers of C. jejuni in the chicken cecum. Amplified DNA fragment concentrations were estimated by agarose gel electrophoresis of the dilution series and subsequent analysis using a Typhoon 8600 variable-mode imager (Amersham Pharmacia Biotech, Little Chalfont, Buckinghamshire, United Kingdom) and the TotalLab Image Master software, version 1.11 (Amersham Pharmacia Biotech). Fragment concentrations were subsequently equilibrated by dilution.

The chain termination and labeling reactions for sequencing were carried out using the primer gltTF and BigDye v.3.1 sequencing chemistry (Applied Biosystems, Foster City, CA) according to instructions supplied by the manufacturer. Sequencing was carried out with an ABI PRISM 3100 genetic analyzer (Applied Biosystems) using a standard protocol.

PCR amplification of 16S rRNA genes was carried out using a universally conserved primer set (17). The PCR mixture contained 1.25 U of AmpliTaq Gold DNA polymerase (Applied Biosystems), 200 μM of a dNTP mixture, 1 μl of template DNA with 0.2 μM of each primer, and 2.5 μM of MgCl₂ in a total volume of 25 μl. The amplification profile was a 5-min activation step at 95°C, followed by 25 cycles of 95°C for 15 s, 50°C for 15 s, and 72°C for 1 min. The reaction was terminated with a 7-min elongation step at 72°C. Fragment concentrations were estimated and equilibrated as described above.

The chain termination and labeling reactions for sequencing were carried out using the universally conserved primer U515F (2) nested within the original fragments and BigDye v.1.1 sequencing chemistry (Applied Biosystems) according to instructions supplied by the manufacturer. Sequencing was carried out as described above.

gltA data. The test set data used for evaluation of the C. jejuni quantification model consisted of 18 mixtures of seven closely related strains of pathogenic Campylobacter jejuni designated G10, G12, G96, G109, G114, G1125, and G147 (Fig. 2, black graph lines). The genetic region used for the analysis was an approximately 230-nucleotide fragment of the glt gene (corresponding to positions 1664595 to 1664823 in the C. jejuni subsp. jejuni NCTC 11168 strain [20]), which has 14 sites that are polymorphic between the strains. Since the data region is relatively poor in information, the modeling tends to be heavily influenced by random variations in the spectra. In order to reduce the variations in the spectral matrix, the peaks for polymorphic sites were culled, i.e., spectral values for the point of base calling as well as three flanking values on each side (for a total of seven measurements per polymorphic site), were extracted from the raw data files, producing 98-by-4 spectral matrices. The same was done for the seven pure strain sequence spectra. Each reduced spectrum was subsequently rescaled by multiplying all values within blocks of 7 by 4 (i.e., emission readings for one polymorphic site) with the ratio between the total spectral mean and the block mean, giving equal weight to each block of variables. Furthermore, each such matrix was unfolded (i.e., column 2 is concatenated onto the end of column 1, column 3 onto this vector, etc.) and mean normalized (i.e., each vector entry is divided by the total vector mean), producing vectors of 392 measurement values.

The gltA gene spectral data from experimental samples were subjected to the same pretreatment.

16S RNA gene data. The data used for PLSR calibration were constructed from cloned 16S rRNA partial gene sequences from a chicken cecal library. The seven GCUs included in the experimental design were representatives of the genera Campylobacter, Clostridium, Escherichia (E. coli), Eubacterium, Faecalibacterium, Lactobacillus, and Ruminococcus (genus names were found by performing a BLAST search of the GenBank database). Eighty-seven training samples of amplified DNA fragments from the seven GCUs were mixed according to a simplex lattice design (4) in which the mixture ratios are spread evenly over the response region. In our case, this means that in the calibration samples, each component varied between 0 and 100% relative abundance (100% being the samples with only one genotype), corresponding to an 87-by-7 mixture matrix, including three central samples where all seven GCUs were mixed in equal ratios. These samples were sequenced, and spectral data corresponding to a 35-nucleotide region (positions 622 to 656 in the E. coli 16S rRNA gene), between the points of base calling for two conserved flanking residues, were extracted from the raw data files for use in subsequent numerical analysis. This region has 20 sites that are polymorphic between the seven GCUs. Even though the extracted data region was identical for each spectrum in terms of the number of nucleotides scanned, the number of emission readings varied slightly from sample to sample. The main cause of this heterogeneity is probably the highly polymorphic nature of the DNA sequences which causes slight differences in migration rates during sequencing electrophoresis. In order to remove this variation, each spectrum was subjected to a cubic spline interpolation technique. This treatment allows for the coercion of all spectra to a given length, which in this case was chosen to be 450 entries per spectral vector. Interpolation of
emission spectra was carried out using the R function “spline” as implemented in the library “stats” using the “fmm” spline-fitting algorithm (6), producing 87 450-by-4 emission matrices. These matrices were unfolded, and mean normalized, and all pretreated spectra were collected in an 87-by-1,800 matrix, i.e., each of the 87 calibration samples was represented by a spectral row vector of 1,800 numeric entries.

The 16S rRNA data used as a test set consisted of a 45-point simplex lattice design with the GCUs *Bacteroides*, *E. coli*, and *Faecalibacterium* (i.e., 45 samples in which the three GCUs are mixed to spread evenly over a response region between 0 and 100% relative abundance), representing a subset of the training data, as well as 7 samples of all seven GCUs mixed in various ratios, for a total of 52 samples (see Fig. 5, black graph lines). The genetic region represented in the raw data was identical to the one described above, and all spectra were given the same pretreatment as that of the calibration samples. The data set was used for testing the prediction accuracy of MLR and PLSR.

Experimental 16S rRNA data were also given the pretreatment described above.

**Model systems for direct sequencing application.** Two model systems were chosen for demonstrating the application of our approach. In the first system, we were interested in the colonization dynamics of *C. jejuni* in chickens. *C. jejuni* is a very important foodborne pathogen with the main reservoir in the chicken gut (28). Understanding how this bacterium colonizes chickens is of crucial importance in preventing it from entering the human food chain. The second model system was designed to obtain basic knowledge about microbial population dynamics in gut systems. Both data sets presented are part of two larger ongoing projects where the aims are to elucidate the biological questions described above.

In the first model system, chickens were infected with a cocktail of the seven *C. jejuni* strains at day 1. Bacterial samples were obtained from cecal scrapings at days 16 (six samples), 20 (five samples), and 23 (six samples). Total bacterial DNA was isolated according to an optimized and automated protocol (27). PCR amplification and sequencing of the ghaA gene were carried out as described above.

The second model system consisted of a growth chamber containing ~1.5 liters of anaerobic basal growth medium (Oxoid, Basingstoke, Hampshire, United Kingdom), continually flushed with a mixture of N2 (80%) and CO2 (20%) in order to achieve anaerobic conditions. This simple fermentor was inoculated with a mixed culture containing the three bacterial species *Bacteroides uniformis* (ATCC 8492), *Escherichia coli* (ATCC 25922), and *Lactobacillus salivarius* (ATCC 11741). These species are identical to the GCUs used for multivariate modeling, with respect to the relevant 16S rRNA gene and primer sequences. The first sample was collected 2 h after inoculation, and subsequent samples were collected hourly for a total of eight samples. Optical density at 600 nm (OD600) and pH were measured at each sampling point. Total DNA was isolated according to the protocol used for the first model system. PCR amplification and sequencing of 16S rRNA genes were carried out as described above.

**Multivariate regression analysis of DNA electropherograms.** The simplest multivariate analytical approach to GCU quantification is MLR using the linear mixture model. This method depends on prior knowledge of the unit spectra, i.e., the pure sequence spectra of the GCUs involved in the mixtures. If the quantitative response of the sequencing apparatus is additive, i.e., a mixture of two or more DNA fragments produces a spectrum which is a linear superimposition of the unit spectra, the mixture can be modeled as a linear combination of the unit spectra of the form

\[ y = b_0 + b_1 x_1 + \ldots + b_i x_i + \varepsilon, \]

where \( y \) is the mixture spectrum, \( x_1, \ldots, x_i \) are the pure unit spectra, \( b_1, \ldots, b_i \) are regression coefficients, \( b_0 \) is the intercept, and \( \varepsilon \) is an error term. The least-squares MLR coefficients \( b_1, \ldots, b_i \) may then be interpreted as the relative contribution of each pure spectrum to the mixture and, thus, as the relative abundances of the different genotypes.

PLSR does not depend strictly on prior knowledge of the unit spectra of the
individual GCUs in a mixture. What is required, however, is a set of calibration spectra corresponding to samples of known mixture ratios. The algorithm is based on the decomposition, \( X = TP + E \), where \( X \) is a matrix containing the calibration spectra, \( T \) is a matrix of scores (latent variables), \( P \) is a matrix of loadings (rotation vectors), and \( E \) is a matrix of residuals. The latent variables in \( T \) are orthogonal, effectively solving the problem of near colinearity, which is common in spectral data, and the extraction of latent variables greatly reduces the dimensionality of the data. Based on the decomposition matrices and the matrix, \( Y \), of known GCU mixture ratios, \( Y \) can be formulated as a linear function of \( X \), such that \( Y = XB + E \), where \( B \) is a matrix of regression coefficients and \( E \) is a matrix of error terms. The PLSR algorithm actively uses information in \( Y \) in order to extract latent variables from \( X \) that are optimal for prediction, making this an appropriate regression method for our purpose. For a detailed discussion of these multivariate regression techniques, we refer to Martens and Næs (13). A more thorough description of the statistical analysis carried out in this work (including experimental design and data preprocessing) can also be found in the supplemental material.

All computations were carried out using statistics programming interface R (R Development Core Team [2006]). R is a language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org). PLSR was carried out using R "pls" software, with the orthogonal score algorithm for fitting multiple response models.

**RESULTS**

**Direct MLR quantification.** An MLR model was evaluated for exploring the colonization kinetics of the seven *C. jejuni* strains in the chicken cecum, using the *gltA* gene as a marker for the seven strains. The 18 mixed *gltA* sequence spectra from the *C. jejuni* test set were modeled as linear combinations of the seven pure strain unit spectra, omitting the offset term, as shown in equation 1, as follows:

\[
y_j = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_6 x_6 + \beta_7 x_7 + \varepsilon_j
\]

(1)

where \( y_j \) (\( j = 1, \ldots, 18 \)) is a 392-entry column vector representing a mixture spectrum, \( x_1, \ldots, x_7 \) are unit spectra of the same length, \( \beta_1, \ldots, \beta_7 \) are regression coefficients, and \( \varepsilon_j \) is an error term. The regression coefficients were converted to percentages and compared to the original mixture ratios (Table 1 and Fig. 2). The graphs show a good match between predicted and measured values, with an overall explained variance of 87.4% and a root mean squared error of prediction (RMSEP) of 5.38. These estimates, however, do show that the assumption of additivity, on which the linear mixture model is based, does not exactly hold. Most prominently, the estimates of strain G109 proportions are systematically biased toward underprediction. This is reflected in the proportion of explained variance, which is greatly reduced relative to the other GCUs, as well as the inflated standard error of prediction (Table 1). Deviation from additivity is also evident in the predicted values for the GCUs G10 and G125, which are generally overpredicted. Correlations between predicted and measured proportions show that the predictions are consistent. Since prediction errors are generally systematic, bias correction may be applied to the proportion estimates. If estimates for G10, G109, and G125 are adjusted by the addition or subtraction of their respective mean bias factors, test statistics are significantly improved, with the total explained variance increasing to 93.5% and the mean prediction error decreasing to 3.86. These results suggested sufficient prediction accuracy for GCU quantification with experimental samples of unknown composition. Although the bias is quite clear and systematic in the present study, further validation is needed before this kind of bias correction can be carried out on a regular basis.

To illustrate the usefulness of our approach, we applied direct sequencing and MLR quantification using the model described above (equation 1) to a time series experiment. The graphs in Fig. 3 show proportion estimates of the seven *C. jejuni* strains in 17 cecal samples from the chicken infection experiment. Regression coefficients \( \beta_1, \ldots, \beta_7 \) were computed for the new mixture spectra, \( y_j \) (\( j = 1, \ldots, 17 \)), and converted to percentages. Bias correction was applied to proportion estimates for G10, G109, and G125 using the mean bias estimates computed from the test set. The analysis shows that the dominant infective strain is G109, which occurs in significant proportions in every individual examined at each time point. There are also sporadic occurrences of strains G114 and G125, with decreasing proportions of G114 from day 16 to day 23 and the opposite trend for G125. The four remaining strains do not show a significant occurrence during the course of the infection.

Using the 16S rRNA gene as a marker for differentiating between GCUs enables us to analyze communities of bacteria belonging to distantly related lineages (19). This is certainly the case for bacteria of the gastrointestinal tract, which display significant phylogenetic diversity (40). The seven GCUs in the 16S rRNA gene data sets were selected to exemplify this diversity as it pertains to the chicken cecum.

The direct MLR quantification approach was applied to the 16S rRNA test set. The 52 spectra in this set were modeled as linear combinations of the seven pure GCU unit spectra from the calibration set according to a model structurally identical to equation 1. In this case, the total explained variance was 94.1%, and the overall RMSEP was 5.61. While these are good results, the proportion predictions were less consistent than those for the *C. jejuni* data, and prediction errors were less systematic (see Fig. SA1 and Table SA1 in the supplemental material).

The 16S rRNA gene data consist of greatly divergent DNA sequences. This makes the electropherograms quite complex, with occasional phase shifts, causing corresponding wave tops for different subsequences to fall out of alignment. In the *gltA* gene data, DNA sequences are very similar, with only 14 SNPs...
scattered over ~230 nucleotides of DNA sequence, greatly reducing the occurrence of phase shifts. This facilitates the culling of a fixed number of scans for polymorphic sites, making each such site a clearly defined data region, which again allows for rescaling over each spectrum. Such a selection of variables was not feasible for the 16S rRNA gene data, and the complexity of these spectra may aggravate the effects of any nonlinearities, making MLR less appropriate for quantification purposes.

Direct PLSR quantification. By using PLSR to calibrate the sequencing apparatus, nonlinearities are taken into account. In order to compute PLSR predictor functions, the 87-by-7 matrix of mixture ratios from the 16S rRNA gene calibration data were modeled as a function of the 87-by-1,800 matrix of pretreated training spectra, using full cross-validation. The model converged on six partial least-squares (PLS) components (Fig. 4) and 96.0% and 96.9% explained $y$ variance for the cross-validation and calibration, respectively. Overall, the RMSEP for the cross-validation was 4.33. The analysis produced seven predictor functions, one for each GCU included in the calibration, as shown in equation 2, as follows:

$$y_i = \beta_{i1}x_1 + \ldots + \beta_{i1,800}x_{1,800}$$  \hspace{1cm} (2)

FIG. 3. Proportion estimates of *C. jejuni* strains for 17 cecal samples from an infection experiment. The first six samples (starting at tick label 16) were collected from six different chickens at day 16 of the experiment. The next five samples were collected at day 20, and the last six samples were collected at day 23.

FIG. 4. $R^2$ (black-line) and RMSEP (red-line) curves for the PLSR cross-validation. The squared Pearson correlations are between predicted and measured $y$ values. As more components are included in the model, RMSEP decreases and $R^2$ values increase precipitously until the model converges on six PLS components.
where $y_i$ is the relative abundance of GCU $i$ ($i = 1, \ldots, 7$), $\beta_1, \ldots, \beta_{1,800}$ are the PLS regression coefficients for the prediction of GCU $i$, and $x_1, \ldots, x_{1,800}$ are the entries of a relevant spectral row vector.

The model was applied to the spectra from the 16S rRNA gene test set, giving GCU proportion estimates for the 52 samples. The total explained $y$ variance for the test set was 96.1%, and the overall RMSEP was 4.56. Figure 5 shows predicted versus measured proportions demonstrating a good match. Prediction statistics are summarized in Table 2. Note that for the groups that were not present in the 45-point simplex lattice design included in the 16S rRNA test set, the explained $y$ variances are heavily influenced by the low expectation values of $y$. Since the expectation is, in most cases, close to that of the measured values, the explained variances for these groups are disproportionately low considering their RMSEPs.

Figure 6 shows the direct sequencing and PLSR quantification approach using the previously computed model (equation 2) applied to samples taken from the anaerobic fermentation experiment. The graphs show that *E. coli* dominates the culture from the 2-h point until the end of the time series. This is not unexpected since *E. coli* is a fast-growing and metabolically versatile bacterium, probably better equipped for handling the transition from the starter culture to the growth chamber than the other two GCUs. It is probable that the initial growth of *B. uniformis*, a strictly anaerobic bacterium, was hampered by exposure to oxygen during preparation of the starter culture, as
this would be the most likely cause for the cells to alter their expression pattern from growth to stress response. Such a response has not been described explicitly for *B. uniformis*, but it has been documented for its relative, *Bacteroides fragilis* (23, 24). The proportion curve of *B. uniformis* shows a declining trend from the starter culture until 8 h, presumably the approximate time required to reset its expression pattern, after which some growth appears to begin. The OD_{600} curve suggests that general log phase growth commences after ~4 h. This coincides roughly with an upsurge in *L. salivarius* proportions and a precipitous drop in pH, which is to be expected due to lactic acid production, which is characteristic of this GCU.

**DISCUSSION**

**Field of application.** In this report we have presented a methodological framework which enabled us to effectively monitor the dynamics in bacterial communities from two different experimental settings. In general, the methods presented above should be applicable to any defined community of organisms within a reasonable range of complexity. One prerequisite for the approach to be effective is the existence of a genetic marker conserved throughout the community or subcommunity of interest, which has a sufficient degree of polymorphism for distinguishing between all the relevant GCUs. Such a marker can then be amplified using a single primer set and subsequently sequenced directly with one common labeling primer. A second prerequisite is the availability of the pertinent pure spectra of the different genotypes in the total population (MLR method) or of a calibration data set of mixtures of known composition (PLSR method). If these conditions are met, proportion estimates for the various GCUs in a community sample can be computed in a relatively straightforward way by multivariate analysis of the community sequence spectrum. Automated sequencing is by now a very robust and economical procedure, making our approach an effective means of collecting time series data from model microbial communities.

It should be noted that our approach is not developed for the simultaneous analysis of all members in communities with high species richness. One problem likely to be encountered when applying the described method with high-richness communities is the presence of rare community members. Our test set data contained GCUs representing as little as 3.6% of the total abundance, and these are predicted with good accuracy. However, considering that this number is of the same order as that of the estimated prediction errors (<5%), relative abundance estimates for rare GCUs should be interpreted with care. A second limitation of our approach pertaining to the analysis of complex communities is the availability of data for modeling. As species richness increases, so does the requirement for input data, which again depends upon a priori knowledge of the system in question. As such, other techniques are more suitable for exploring species-rich natural communities as a whole. Nevertheless, as we have demonstrated, our approach can be quite effective when the objective is to monitor subgroups within complex communities.

There are also certain challenges that are intrinsic to any PCR-based protocol for microbial community analysis (36). One potential source of quantitative bias is preferential DNA extraction. The choice of DNA isolation method has been shown to influence the outcome of quantitative analyses of
environmental samples (11, 14), and care should be taken by investigators to minimize this source of error. In our assays, we used a very harsh mechanical treatment in order to achieve unbiased cell lysis (27). A second source of bias concerns differential amplification efficiency in mixed template PCRs. Amplification bias could be introduced by factors such as differing efficiencies in primer-template hybridization (21), differences in genome sizes and rRNA gene copy numbers (5), changing stoichiometrics during the course of the reaction (32), and stochastic effects (21). Again, the individual experimenter needs to take precautions both when performing the experiment and when interpreting the results. In this work, we used only short amplicons in order to overcome some of the challenges associated with amplification bias in mixed template PCRs. Furthermore, we used a specially selected 16S rRNA gene amplicon, the quantitative properties of which have been fairly well documented (17, 25).

It should be noted that as far as the above-mentioned bias effects are systematic and reproducible, they should have less bearing on quantitative results when the objective of the study is delineating dynamics. Even if community structure at some time point is skewed by the experimental procedure, a change in structure from that point to the next should still reflect actual community dynamics.

**MLR versus PLSR for GCU quantification in mixed samples.** The overall performance of MLR using the linear mixture model for relative GCU abundance estimation suggests a fairly good predictive ability. One obvious advantage with this approach is the minimal need for information input into the modeling. Significant prediction bias can be assessed using a relatively small test set and may be taken into account when results are interpreted, as illustrated by the analysis of the C. jejuni data. This method, however, does turn out to be somewhat less consistent when the genetic data represent highly polymorphic DNA sequences, as for the 16S rRNA gene. In such a situation, MLR is probably best suited when only rough estimates are needed.

For the analysis of the more complex 16S rRNA gene data, a PLSR calibration approach proved more effective than MLR. The pretreated 16S rRNA electropherograms contain a number of variables (1,800 X variables when used for PLSR) that is much larger than the number of samples used for calibration (87 samples). In this type of data, near collinearity is often a prominent feature, making ordinary least-squares regression coefficients unstable (18). The PLSR algorithm effectively removes this problem by finding new sets of orthogonal latent variables that are linear combinations of the original predictor variables. By including only latent variables that contribute significantly to the model's explanatory power and discarding the remainder as noise, this algorithm can provide robust predictors (13). Our PLSR model converged on six PLS components. Since we were dealing with a system of seven GCUs, their proportions may vary with only 6 degrees of freedom. Thus, our model reflects the underlying dimensionality of the system. This indicates that very little noise was incorporated into the model, making our linear predictors for GCU quantification stable. It should also be noted that in general, the algorithms presented here should perform even better if the mixture data used for calibration and validation are free from errors due to pipetting and inaccuracies in estimation of DNA concentrations.

**Comparison with existing technology.** Several existing technologies are well suited for investigating community structure at various levels of detail. Community fingerprinting techniques are relatively efficient, both in terms of cost and labor, and have been used extensively (9, 39, 41). The main limitation of fingerprinting methods lies in quantification, as they tend to produce patterns rather than information about specific taxa. Nevertheless, fluorescence-based fingerprinting methods can provide useful quantitative information on microbial community structure (10). Large-scale metagenomic analyses and similar environmental sequencing projects have, in several cases, provided detailed pictures of taxonomic and metabolic diversity with microbial community samples (1, 7, 29, 35). However, due to both the cost and workload associated with this type of analysis, it is currently not practical for providing high temporal or spatial resolution. Community microarrays show some promise as tools in microbial ecology (38), but working with microarrays is both relatively expensive and laborious. Furthermore, standard hybridization of array-bound oligonucleotide probes to target sequences has been shown to be poorly explained by thermodynamic parameters, making construction of quantitative community microarrays quite difficult (22). A recent development in facilitating microarray design is probe labeling according to the microsequencing principle (33). The associated workload, however, is still large compared to that of the approach described here.

**Concluding remarks.** For the various reasons stated above, current techniques are, in general, not feasible alternatives for generating quantitative community data of the temporal resolution needed for deciphering microbial community dynamics and interactions. Like animals and plants, bacteria exist in communities in which interactions through phenomena, such as competition, density dependence, predation, and cooperation, occur within and among populations (15, 16, 31, 34). These are fundamental properties of any ecosystem, yet very little is known about population dynamics in microcosms. This lacuna in the existing literature is most likely related to the absence of convenient data acquisition methodologies, as well as to a historical division between the scientific disciplines of microbiology and traditional ecology. We believe that our approach, facilitating the collection of microbial community time series data, may contribute significantly to bringing microbial ecology up to speed. Moreover, microbes as model systems offer some great advantages relative to larger organisms (8). Their short generation times allow for the monitoring of potentially hundreds of generations within a week, while their small sizes facilitate easily manipulable laboratory settings. As such, microbes can provide a convenient basis for addressing more general problems in ecology and evolution. A more profound understanding of microbial communities will also be of great value within fields such as medicine, epidemiology, and environmental science.

**ACKNOWLEDGMENTS**

We thank Signe Marie Drømtorp for excellent technical assistance. This work was financially supported by a research levy on certain agricultural products and by Hedmark Sparebank.
REFERENCES


