Improved Assessment of Denitrifying, N₂-Fixing, and Total-Community Bacteria by Terminal Restriction Fragment Length Polymorphism Analysis Using Multiple Restriction Enzymes

Christopher Rösch and Hermann Bothe*
Botanisches Institut, Universität zu Köln, Cologne, Germany

Received 26 August 2004/Accepted 10 November 2004

A database of terminal restriction fragments (tRFs) of the 16S rRNA gene was set up utilizing 13 restriction enzymes and 17,327 GenBank sequences. A computer program, termed TReFID, was developed to allow identification of any of these 17,327 sequences by means of polygons generated from the specific tRFs of each bacterium. The TReFID program complements and exceeds in its data content the Web-based phylogenetic assignment tool recently described by A. D. Kent, D. J. Smith, B. J. Benson, and E. W. Triplett (Appl. Environ. Microbiol. 69:6768-6766, 2003). The method to identify bacteria is different, as is the region of the 16S rRNA gene employed in the present program. For the present communication the software of the tRF profiles has also been extended to allow screening for genes coding for N₂ fixation (nifH) and denitrification (nosZ) in any bacterium or environmental sample. A number of controls were performed to test the reliability of the TReFID program. Furthermore, the TReFID program has been shown to permit the analysis of the bacterial population structure of bacteria by means of their 16S rRNA, nifH, and nosZ gene content in an environmental habitat, as exemplified for a sample from a forest soil. The use of the TReFID program reveals that noncultured denitrifying and dinitrogen-fixing bacteria might play a more dominant role in soils than believed hitherto.

The characterization of bacterial communities in environments is complicated by extreme species richness and a high variability of distribution within short distances. Soils are estimated to harbor up to 10¹⁰ bacteria of about 10⁴ different ribotypes per g (3, 19), of which more than 95% cannot be cultured by present methods (2, 11). However, knowledge about microbial community structure and its variations is required to predict nutrient fluxes and to analyze the effects of xenobiotics. In a more recent avenue, molecular methods have been introduced to get access to the noncultured population of bacteria in environmental samples. Most approaches published so far either rely on DNA sequencing of clone libraries (16, 21) or employ fingerprinting techniques (for example, denaturing gradient gel electrophoresis [DGGE] [12] or terminal restriction fragment length polymorphism [tRFLP] analysis [9]) to demonstrate microbial diversity or population shifts in microbial communities. Clone libraries supply novel sequence information and allow the phylogenetic identification of individual clones, but this approach is laborious and expensive. PCR-based fingerprinting techniques such as DGGE or tRFLP (9) primarily provide population-specific signatures. DGGE bands can be sequenced but usually yield relatively short sequences. Although tRFLP bands cannot be sequenced, this method is preferred to DGGE, because tRFLP allows analysis of population structures of more complex communities. However, since individual peaks of a tRFLP profile determined using a single restriction enzyme can be derived from a whole range of nonrelated species in many cases, tRFLP is used to characterize shifts in a population structure but not to identify individual strains in a community. To circumvent this limitation, a method has recently been introduced to produce a species list for a bacterial population of unknown composition on the basis of experimental tRFLP data from separate digests of multiple different restriction enzymes (6). This phylogenetic tool is based on analysis of fluorescence-labeled terminal restriction fragments of 16S rRNA gene segments. Independently of that work (6), this laboratory generated software by a similar approach that also employed the 16S rRNA gene; however, this approach uses a different sequence area which is present in most sequences deposited in the databanks. For this computer program, termed TReFID, data from up to 13 different restriction enzymes were used and a unique algorithm was developed. This computer program allows a statistical analysis of experimental data and comparison with reference sequences from public databases.

In addition to this phylogenetic assignment method for the analysis of tRFLP profiles of the 16S rRNA gene, the present work describes, for the first time, the development of a similar tool for N₂ fixation and denitrification. For N₂ fixation, nifH coding for dinitrogen reductase was chosen because this gene is highly conserved among microorganisms (22) and because some 2,000 sequences are deposited in GenBank. For denitrification, nosZ, encoding nitrous oxide reductase, with only about 180 entries in GenBank, was selected. A further disadvantage of this gene is that it does not occur in denitrifying bacteria which do not reduce N₂O. However, this gene had to be chosen because all of the other steps of denitrification (reduction of nitrate to nitrite, nitrite to nitric oxide, and nitric oxide to nitrous oxide) are catalyzed by at least two different enzymes in each case. Moreover, the sequence information for all these genes is even scarcer than that available for nitrous oxide reductase.

To demonstrate the potential of the present approach, the
bacterial population of a forest soil from the vicinity of Cologne (Germany) was analyzed for its content with respect to the 16S rRNA, nifH, and nosZ genes. For this goal, a soil was selected with a high carbon/nitrogen ratio, which led us to assume that both denitrifying and N₂-fixing bacteria occurred there simultaneously and with a high abundance. DNA was extracted from samples of the soil, and segments of the three genes were amplified by PCR and subjected to tRFLP analysis. Additionally, PCR products generated from the DNA of the forest soil were cloned and sequenced. The species list obtained by the TReFID software was compared with the sequence data from the clone libraries. Data were also evaluated by various control procedures which will be shown in detail. It is suggested that the tRFLP analysis employing multiple restriction enzymes will become a useful tool to analyze complex microbial communities and one which will improve as more sequence information for the different enzymes becomes available.

MATERIALS AND METHODS

Soil samples used and techniques for the molecular characterization of the DNA. The soil analyzed for its bacterial content came from Dünwald forest (51°10′50″N, 7°03′17″E) in the suburban area of Cologne, Germany. This loamy sand soil had a pH of 3.9 and a content of water of about 10% (wt/wt) and of humus of 4 to 6% (wt/wt) on the sampling date of 2 October 2002. The high C/N ratio of 18.3 to 19.1 was particularly noteworthy. The total N content was 0.28 to 1.48 g/kg (wt/wt) and the humus of 4 to 6% (wt/wt) on the sampling date of 2 October 2002. The high C/N ratio was 65 to 50°C in the case of amplifying the 16S rRNA, whereas some minor modifications were made for improvements. PCR products were obtained in GeneScan 3.1.2 and verified by BlastN (1), ClustalX alignments (18), and ChimeraCheck (8). The phylogenetic affiliation of the novel clones was deduced by means of the RDP II Phylogeny Interface (http://rdp.cme.msu.edu/alg/phylop.cgi). Additionally, neighboring joining phylogenies of 100 replicates trees were constructed with ClustalX and visualized with TreeView 1.6.1 (14). Gap positions were excluded from the analysis, but corrections for multiple substitutions were applied. Corrected sequences were deposited in GenBank (www.ncbi.nlm.nih.gov [accession no. AY723961 to AY724250]).

For tRFLP analyses, 5’ fluorochrome-labeled PCR primers were used: 63F-6-carboxyfluorescein or 63F-6-carboxy-4,5-dimethoxyfluorescein (JOE) for the 16S rRNA gene and nifHF-6-carboxyfluorescein, or nosZR-6-carboxytetramethylrhodamine (MWG Biotech, Ebersberg, Germany). Purified PCR products from a reaction with nonlabeled primers were reamplified in a second PCR with labeled primers up to a total volume of 1,000 μl of PCR product. Products of this second reaction were purified (QiAQuick gel extraction kit [QIAGEN] or Ultrafree-MC columns [Millipore, Bedford, Mass.]), and in 2012 also showed that PCR products had been obtained in sufficient quantities. The preparation was then partitioned into up to 13 aliquots and digested overnight at 37°C in a 100-μl volume by separately using one of the following restriction endonucleases (MBI Fermentas, Leon-Rot, Germany) per tube: Alul (AC/CT), Bme1390I (CC/CGO), Bsh1236I (CO/CG), Clrl3I (G/GNC), HaeIII (GG/ CC), HinfI (CG/C/G), HinfII (G/A/TC), MboI (G/ATC), MspI (CC/GG), or Rsal (GC/AC). Digestions with Tail (AAGCT), Tacl (T/CA/G), and Tacl (AATT) were performed at 65°C. After ethanol precipitation, the fragment mixtures were dissolved in 10 μl of sterile deionized water. Prior to gel loading, the samples (2.5 μl) were mixed with formamide (1.0 μl), loading buffer (Applied Biosystems) (1.0 μl), and a GeneScan 500 ROX size standard (Applied Biosystems) (0.5 μl). The analysis of the 1.6 μl of this mixture was performed by obtaining a 36-cm-long 4.5% polyacrylamide gel at 2,700 V (ABI 377 automatic sequencer equipped with GeneScan 3.1.2; Applied Biosystems). The sizes of fragments were determined by using the local Southern method implemented in GeneScan 3.1.2 and a GS-500 ROX size standard. Only fragment lengths in the range of 30 to 500 nucleotides (nt) were considered for analysis. The noise threshold for signal processing was set as low as possible (generally about 20 relative fluorescence units) to cover a broad range of prominent and weak peaks at the same time. The peak heights for single tRFLP profiles were distributed over 2 orders of magnitude (~20 to 200 relative fluorescence units). The results were exported as tabulated-delimited text files (GeneScan 3.1.2 and further processed with TReFID (Fig. 1).

Software development. The source code of the data analysis software, TReFID, was written in PureBasic 3.80 and compiled to FASM assembler language. TReFID is a small stand-alone executable file for Microsoft Windows XP. It was obtained as a ZIP file from http://www.trefid.net and can be installed by double-clicking on the setup file. The software uses a MySQL database and a Microsoft SQL Server database as the database repository. A graphical interface provides access to most of the software functions. TReFID employs a MySQL database to store the current data and the results of analysis. The database is selected at the program start. These data are sorted and displayed in three different ways. In a first-tree-view control, all species contained in the selected database are displayed in alphabetical order. A second tree-view control summarizes all entries according to their phylogenies (based on Bergey’s Manual 2001). The third two-view control arranges the entries by their expected terminal restriction fragments (tRFs) for each restriction enzyme in use. As examples, databases are provided for the 16S rRNA (63F), nifHF (nifH), and nosZ (nosZR) genes. Currently, the only operating system supported is Microsoft Windows XP.

Preparation of the databases. DNA sequence data for the three genes examined (the 16S rRNA, nifH, and nosZ genes) were downloaded in GenBank format by use of the Entrez nucleotide database query form (www.ncbi.nlm.nih.gov), and all available sequences (October 2003) were analyzed using multiple sequence alignments (ClustalX 1.81) (18). Sequences including the binding sites for the fluorochrome-labeled primers (nifHF and nosZ) were incorporated in the respective TReFID databases. These sequences had to be modified to fit the 5’ end of the primer for nifHF or the 3’ end of that for nosZR. Overhang nucleotides were cut, while missing nucleotides (up to 30) in the primer binding site were completed with “N” to assure correct fragment grouping. This modification was automatically accomplished by use of our own GBSD program (available under http://www.trefid.net) and a table of the number of nucleotides (up to 30) to be removed or added to the reference sequences. Alignments of all sequences were manually done with ClustalX. Only 16S rRNA gene sequences deposited before January 2003 were used and similarly analyzed. A single GenBank sequence can be characterized by its set of theoretically derived tRFs. A graphical representation of this would be a polygon in a spider web graph (Fig. 2), where each axis represents a different restriction enzyme; the respective fragment sizes are dots on these axes, the origin being in the center.

Data analysis. To produce a species list with the TReFID software, both the desired database and the experimental tRFLP data have to be loaded. Two analysis parameters can be set by the user to allow fine tuning. (i) The lengths of tRFs determined mathematically can differ from those of the experimentally determined tRFs obtained from the DNA of an environmental sample (see Results). Therefore, the latter are not treated as discrete values by TReFID but as intervals with score values of 1.0, 0.5, or 0.25. (ii) The tRFs obtained from the DNA of an environmental sample have to match the tRF polygon of a sequence deposited in the TReFID database bank. The minimum number of tRFs which have
to be retrieved from a polygon is two of three (i.e., 8 enzymes with matches out of total 12 employed, for example). This is defined as the threshold value (see Results). Enzymes with expected fragments more than 500 nt in length were excluded from the analysis of a sequence. Predicted tRFs less than 30 nt in length were allowed when a restriction site was located inside the primer binding site. In this case, the second restriction site, located outside the primer sequence, was used for the TReFID database and also for the analysis of the sample DNA. Those tRFs of a DNA to be analyzed in cases in which

**Polygon.** A graph is formed by all tRFs of one GenBank entry (thus for one organism or one sequence deposited) by use of all restriction enzymes (Fig. 2). For evaluation of the polygons, only those formed by the tRFs with a threshold of ≥2/3 are taken for analysis of the polygons deposited in the databank.

**Match.** tRF is retrieved from the TReFID databank, with a deviation between expected and experimentally obtained lengths of not more than 1.5%.

**Score.** Deviations between determined and theoretical tRF lengths as found in the TReFID bank are scored as follows: results of ≤0.5% receive a score of 1.0; results between 0.5 and 1.5% receive a score of 0.5; results between 1.0 and 1.5% receive a score of 0.25; and results of >1.5% receive a score of 0.

**Score sum.** The sum for the tRFs with all the restriction enzymes employed referred to a single database entry (a polygon for an organism).

**Threshold.** For the threshold value, ≥2/3 of restriction enzymes have to provide a tRF with a score of >0 for an analyzed sequence.

TReFID. TReFID represents the terminal restriction fragment identifying program.

**RESULTS**

Content of the tRFs for *nifH*, the 16S rRNA gene, and *nosZ* in the TReFID databank. The terminal restriction fragment identifying program, termed TReFID (see Materials and Methods) (Fig. 1) was developed to retrieve the tRF pattern of a specific organism for each of the three *nifH*, 16S rRNA, and *nosZ* genes within the total tRF data deposited in the TReFID bank. For this, in the case of *nifH* 1,873 GenBank sequences available in October 2003 have been examined with respect to their *nifHF* primer binding site content. Of these, 1,031 did not contain this site and 20 did not show sufficient homology for primer binding. Primers constructed from other *nifH* regions would not have provided more sequences (unpublished data).

Thus, the resulting 822 sequences had to be sufficient for analysis by computer for restriction sites downstream from the *nifH* binding site. This procedure provided for each restriction enzyme a different number of distinguishable, unique tRFs, ranging from a minimum of 30 (with HinfI) to a maximum of 78 (with MspI or Bsh1236I). This is visualized by the black dots in the axes of the spider web graph of Fig. 2I, in which data are presented for the tRFs derived from the GenBank sequences (panel a) and for those experimentally obtained from the DNA isolated from the Dünnwald soil sample (panel b). In the case of *nifH*, only eight restriction enzymes out of the 13 readily available with a 4-bp recognition site were employed. In total, for the eight restriction enzymes, 6,020 tRFs were obtained, of which only 505 were unique (Table 1). This small number is due to the fact that a cut within a conserved motif can give identical tRFs for different organisms.

All 6,020 tRFs were taken for further analysis. Thus, for *nifH*, 822 polygons, each representing an entry for an organism, were derived from the GenBank sequences, and 139 polygons came from our clone library. To demonstrate the feature of a polygon for a specific bacterium, the tRFs for *Azospirillum brasilense* Sp7 are visualized, as represented by the shaded area in Fig. 2I, panel a. All the different polygons can be used to screen for related entities (organisms) with the tRFs provided by an environmental sample.

In the case of the 16S rRNA gene, all 13 restriction enzymes were employed. The sequence information for this gene in

**FIG. 1.** A flow chart describing the steps in the identification procedure of the TReFID program.
GenBank, plus 135 sequences from our Dünnwald clone library, was more extensive than the data for \textit{nifH}. Altogether, 31,927 sequences out of the 73,245 available in January 2003 were analyzed. Among these, 17,462 containing the 63F primer binding site were suitable for the generation of tRFs from the 63F primer (Table 1). This provided 183,902 tRFs altogether, of which only 4,536 were unique. The 17,462 polygons developed for 16S rRNA gene represent a solid basis of data for further analysis.

Unfortunately, the sequence information for \textit{nosZ} is limiting. Only 85 out of 181 GenBank sequences available in October 2003 were suitable for constructing tRFs by use of the nosZR primer which binds near the 3’ terminus of the gene. Together with the 43 sequences of the clone library, these GenBank sequences provided 1,038 tRFs, with only 129 being unique (Table 1). The present meager sequence information for \textit{nosZ} is reflected by the few dots in the corresponding spider web graph (Fig. 2I, panel a) and the low number (128) of polygons obtained. In contrast, the sequence information for the 16S rRNA gene is much more than can be shown in a

FIG. 2. Graphical representation of the tRFs obtained for \textit{nifH} of \textit{N}_2-fixing bacteria and \textit{nosZ} of denitrifying microorganisms. (I) For \textit{nifH}, tRFs derived from the sequences deposited in GenBank (a) and tRFs experimentally obtained from the DNA isolated from the Dünnwald soil (b) are shown. (II) For \textit{nosZ}, tRFs derived from the sequences deposited in GenBank (a) and tRFs experimentally obtained from the DNA isolated from the Dünnwald soil (b) are shown. The shaded polygon symbolizes the tRFs from \textit{Azospirillum brasilense} Sp7. The lengths of the tRFs are given on the axes in the range between 0 and 500 bp.
denitrifying and N2-fixing bacteria which would enable us to
Therefore, we expected a relatively high abundance of both
other locations in the Cologne area which are N saturated (16).
cause this soil was not rich in nitrogen, in contrast to many
The loamy sand soil of the Dünnwald forest was selected be-
method was tested with forest soil from the vicinity of Cologne.
DNA from an environmental sample.

The spider web graph similar to those for nifH and nosZ, because an axis for any restriction enzyme would contain too many dots (tRFs) to be graphically resolved.

**Demonstration of the applicability of the approach by using DNA from an environmental sample.** The applicability of the method was tested with forest soil from the vicinity of Cologne. The loamy sand soil of the Dünnwald forest was selected because this soil was not rich in nitrogen, in contrast to many other locations in the Cologne area which are N saturated (16). Therefore, we expected a relatively high abundance of both denitrifying and N2-fixing bacteria which would enable us to retrieve tRFs of both groups. DNA isolated from this soil was used for PCR amplifications with the fluorochrome-labeled nifHF primer. Restriction digests with the eight different enzymes yielded 284 tRFs in total (Fig. 2I, panel b). These tRFs were then examined for related entities (organisms) with the TReFID program (Table 1). When a tRF for a given restriction enzyme from the sample matched with one of the databank with 0.5% of the total nucleotide length, the score was set to 1.0. When the similarity between the determined length of a Dünnwald tRF and that of the closest one from the TReFID databank differed up to 1%, the score was set to 0.5. When such a difference was even 1.5%, the score was defined as 0.25. Any value outside of this range was defined as 0. Those restriction enzymes which did not provide a tRF for a segment because of the absence of a restriction site within the 30- to 500-bp sequence were not considered any further.

The polygon for *Vibrio diazotrophicus* was selected to illustrate the method. The enzymes AluI and MboI had no restriction site within the 400 bp between the nifHF and nifHR motifs and therefore could not be employed for analysis. Among the tRFs of the residual six restriction enzymes used, five gave the maximal score, meaning that the size of the best matching tRF showed less than 0.5% deviation from the predicted one. However, one enzyme (HinII) gave a fragment of nearly the same size (being maximally 1.5% larger or smaller than that predicted fragment for *V. diazotrophicus*). The score sum for *V. diazotrophicus* in the Dünnwald soil is thus 5.25 out of 6 (88% identity). Thus, the program indicates with high fidelity that an organism occurs in the Dünnwald soil which is closely related to *V. diazotrophicus*.

As another example, in the case of the noncultured clone DUN1+B26, all eight restriction enzymes provided a tRF, but the similarity value was 6.75 out of 8 due to results that included six scores of 1.0, one score of 0.5, and one score of 0.25. The overall similarity between the pool of tRFs from the Dünnwald soil and DUN1+B26 was 6.75/8 (84%), indicating that tRFs of a bacterium closely related to the deposited clone from Dünnwald, DUN1+B26, had been retrieved. The method does not permit one to decide whether this bacterium retrieved by the tRF pattern was identical to that from the clone library. Clearly, the lower the score sum (in percent similarity), the lower is the probability that an organism with sequence similarity to any bacterium of the TReFID bank is present in the environmental sample. A match in only two cases, for example, could have been due to restriction sites at the same sequence position in two totally unrelated bacteria and would therefore be meaningless.

A threshold value was arbitrarily set for all three genes. In the case of the 16S rRNA gene, the tRFs, obtained from at least 9 restriction enzymes out of the 13 maximally used and providing a tRF within the range of 30 to 500 bp, had to give a score of 1.0, 0.5, or 0.25 (and thus matches of at least two of three) to be considered for further analysis (Table 2). Thus, the tRFs of the DNA of an environmental sample also form a polygon. To be used for the further analysis, at least 9 tRFs of the DNA of this sample had to match the 13 tRFs of one polygon present in TReFID. Cases of matches giving results of <2/3 were discarded. If one (or more) restriction enzyme(s) had no site within the PCR fragment analyzed, the tRF for this enzyme was left out. Matches were then referred to 12 (or less) enzymes, but the required threshold value was also kept at ≥2/3.

**Table 1.** Summary of the total sequences and tRFs obtained for the three 16S rRNA, nifH, and nosZ genes

<table>
<thead>
<tr>
<th>No. of restriction enzymes providing a tRF</th>
<th>16S rRNA</th>
<th>nifH</th>
<th>nosZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sequences</td>
<td>Total no. of tRFs</td>
<td>No. of sequences</td>
<td>Total no. of tRFs</td>
</tr>
<tr>
<td>13</td>
<td>1,087</td>
<td>14,131</td>
<td>1,104</td>
</tr>
<tr>
<td>12</td>
<td>3,995</td>
<td>47,940</td>
<td>540</td>
</tr>
<tr>
<td>11</td>
<td>4,715</td>
<td>51,865</td>
<td>228</td>
</tr>
<tr>
<td>10</td>
<td>3,557</td>
<td>35,570</td>
<td>137</td>
</tr>
<tr>
<td>9</td>
<td>2,501</td>
<td>22,509</td>
<td>111</td>
</tr>
<tr>
<td>8</td>
<td>1,002</td>
<td>8,016</td>
<td>57</td>
</tr>
<tr>
<td>7</td>
<td>378</td>
<td>2,646</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>141</td>
<td>846</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
<td>220</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>140</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>17,462</td>
<td>183,902</td>
<td>951</td>
</tr>
</tbody>
</table>

*Number of unique tRFs, 4,536. Number of entries per unique tRF, 40.5.
 Number of unique tRFs, 505. Number of entries per unique tRF, 11.9.
 Number of unique tRFs, 129. Number of entries per unique tRF, 8.0.

**Table 2.** Similarities of the tRFs polygons obtained from DNA of the Dünnwald soil with those of the TReFID databank

<table>
<thead>
<tr>
<th>Similarity (%) calculated score</th>
<th>16S rRNA</th>
<th>nifH</th>
<th>nosZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>94–100</td>
<td>122</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>87–93.9</td>
<td>215</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>80–86.9</td>
<td>640</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>73–79.9</td>
<td>999</td>
<td>47</td>
<td>5</td>
</tr>
<tr>
<td>66–72.9</td>
<td>414</td>
<td>105</td>
<td>7</td>
</tr>
<tr>
<td>59–65.9</td>
<td>0</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>52–58.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>2,390</td>
<td>191</td>
<td>24</td>
</tr>
</tbody>
</table>

*A high percent similarity value indicates that a bacterium occurs in the Dünnwald soil which is closely related to one with its sequence in the TReFID databank. A value of 100% means a perfect match (within &lt;0.5 %) for all restriction enzymes employed. The relatedness (or probability of occurrence) decreases with similarity percentage.*
All the matches giving results of $\geq 2/3$ were further analyzed for their score sum by the TReFID program. An overall similarity value (i.e., the score sum/number of restriction enzymes with a restriction site in the sequence range analyzed) between 94 and 100% was regarded as being highly indicative of similarity to a polygon consisting of tRFs for a specific bacterium of the TReFID databank. In the case of the 16S rRNA gene, the majority of sequences of the TReFID results (1,037) listed for the Dünnwald soil had similarity values between 73 and 79% (Table 2).

In the case of nifH, five out of eight restriction enzymes had to give a match with a score of 1.0, 0.5, or 0.25 (i.e., a match of two out of three at least) for the sequences to be considered any further. Remarkably, the average similarity value was about 10% less in the case of nifH than that for the 16S rRNA gene (Table 2).

The same protocol was tried for nosZ (Fig. 2II). The threshold was set to six out of nine restriction enzymes utilized (i.e., a match of at least two out of three). The highest matches of a polygon of Dünnwald soil DNA were to Pseudomonas stutzeri, with a score sum of 7.0 out of 8 restriction enzymes that had a restriction site inside the fragment analyzed, Paracoccus pantotrophus, with a score sum of 6.0 out of 7, and Azospirillum lipoforum, with a score sum of 6.0 out of 8. Thus, organisms closely related to the denitrifying bacteria mentioned occurred in the Dünnwald soil. The observed similarity value of the Dünnwald sequences for nosZ was as low as that for nifH (Table 2), although the size of the databank for nosZ was too small to allow us to draw a definitive conclusion here.

**Controls to assess the quality of the TReFID program and to ascertain the feasibility of the approach.** Control A. The 16S rRNA gene tRFs obtained from the Dünnwald soil DNA were examined for matches with any of the 135 polygons of the Dünnwald clone library as part of the TReFID databank (Fig. 3a). For this examination, the stringency in the two parameters used to compare the tRFs from the soil DNA and Dünnwald clone library was varied. The ordinate in the figure indicates the percentage of hits of polygons (i.e., the sum of tRFs for each bacterium deposited) obtained from the DNA isolated from the Dünnwald soil (i.e., entity 1) within the total entries of the Dünnwald clone library (i.e., entity 2). The percentage of sequences retrieved was low when the deviation was restricted to $\leq 0.33\%$ and was high at a deviation of $\leq 1.00\%$. For the further calculations, the values 0.66 (for the number of matches) and 0.5% (for the deviation) were selected (dark column in both parts of the figure). However, score values of 1.0, 0.5, or 0.25 (for definitions, see text) were taken for the calculations and this figure.

![FIG. 3. Identification of polygons representing organisms in the Dünnwald soil in two libraries. DNA was isolated from the Dünnwald soil, and the lengths of the tRFs obtained with all 13 restriction enzymes were experimentally determined for the 16S rRNA gene. The lengths of the tRFs were then compared with those calculated using the sequences from the Dünnwald clone library (a) and from GenBank (b). The ordinate represents the percentages of polygons of the soil DNA among the total polygons of either the Dünnwald clone library (a) or the GenBank sequences (b). One abscissa (front side) denotes the proportions of enzymes providing a tRF with a match corresponding to the total number of enzymes employed and providing tRFs for analysis. The other abscissa (right side) denotes the deviations in the lengths of the tRFs. The percentage of sequences retrieved was low when the deviation was restricted to $\leq 0.33\%$ and was high at a deviation of $\leq 1.00\%$. For the further calculations, the values 0.66 (for the number of matches) and 0.5% (for the deviation) were selected (dark column in both parts of the figure). However, score values of 1.0, 0.5, or 0.25 (for definitions, see text) were taken for the calculations and this figure.](http://aem.asm.org/)

![Downloaded from http://aem.asm.org/](http://aem.asm.org/) by guest

**VOL. 71, 2005 IMPROVED ASSESSMENT OF BACTERIAL COMMUNITY COMPONENTS 2031**
entities with a maximal deviation of 0.33% (Fig. 3). Taking all
these difficulties into account, the proportion of sequences
matching between both entities is remarkably high. Clearly,
direct sequencing of the clones obtained provided more accu-
rate data but is much more time consuming and expensive than
TReFID analysis.

The tRFs from the Dünwald soil DNA were also screened
for their occurrence in sequences from the total TReFID da-
tabank, presently containing 17,327 entries (Fig. 3b). Quite a
lot of polygons can apparently be retrieved from this bank.
However, few polygons are exactly the same when developed
from the environmental sample (Dünwald soil DNA) and
compared with those deposited in the TReFID databank (Fig.
3b, front right-side corner column).

From the Dünwald soil DNA, 1,373 polygons formed by
tRFs scoring above the threshold (i.e., ≥2/3 matches) in the
total 17,327 entries of the 16S rRNA gene TReFID databank
could be retrieved altogether. The corresponding values for
the nifH and nosZ sequences were 130 out of 824 and 23 out of
97 total entries, respectively. To demonstrate the specificity
of the TReFID bank, tRFs of the Dünwald soil DNA were
screened in the TReFID bank for the heterologous, false
genes. For nifH tRFs, only three matching 16S rRNA gene
tRFs of the Dünwald soil were obtained by PCR using fluorochrome labeling, and
their nucleotide lengths were determined experimentally. They
corresponded well with those of the reconstructed profile.
However, the peak height was small in some cases and was
slightly above the peak detection threshold value, set at 20
relative fluorescence units (see peaks at 135, 325, or 380 nt in
Fig. 4b). Thus, it was not clear in these cases whether such
small peaks indicated a hit for a TRF. Despite this drawback,
this procedure for constructing a synthetic tRF profile allows
verification of the composition of a bacterial population for
any environmental habitat with respect to the 16S rRNA, nifH,
and nosZ genes or any other gene with a tRF data bank when
this approach is applied to the tRFs obtained for all restriction
enzymes.

Control C. The whole 16S rRNA gene TReFID databank
was examined for the occurrence of organisms with tRF pro-
files (polygons) similar to those of three selected organisms
picked out of the same databank, namely, Escherichia coli
(E05133), Azospirillum brasilense (AY324110), and Rhizobium
leguminosarum (AF5333683). As before, the match in the tRFs
had to be at least two of three to hit a polygon of one of the
three organisms. The TReFID databank provided 537 poly-
gons related to the above-named organisms. Their sequences
available in the TReFID databank showed that only 257 of
them were at least 500 bp in length and thus suitable for
constructing a phylogenetic tree using the neighbor-joining
method. A total of 192 sequences remained after the elimina-
tion of identical sequences deposited with different accession
numbers. These were utilized to construct a phylogenetic tree
for the same region used for predicting the tRFs. The phylo-

FIG. 4. Comparison in the tRF profiles obtained experimentally with DNA from the Dünwald soil and predicted from the sequence
information in the TReFID result list. DNA was isolated from the Dünwald soil probe, and the tRF lengths were determined experimentally. By
using the polygon constructions from the tRFs of all 13 restriction enzymes, a list of bacteria in the Dünwald soil which are closely related to
organisms with entries in the TReFID databank could be compiled. The tRFs of these closest relatives in the databank were then taken to construct
the predicted tRF profile of the DNA from the Dünwald soil.
genetic tree provided a clear separation of the potential organisms belonging to *E. coli*, *R. leguminosarum*, or *A. brasilense* in 183 out of 192 cases (Fig. 5). However, sequences of three bacteria (*Pseudoalteromonas prydzensis*, *Shewanella gelidimarina*, and *Idiomarina baltica*) (plus three closely related sequences which did not separate in the phylogenetic tree), as well as sequences from three noncultured proteobacteria, did not fit at all. Therefore, these artifacts were analyzed in more detail. In the case of *S. gelidimarina* (AF530149), the eight enzymes AluI, Bsh1236I, Cfr13I, HaeIII, HinfI, MboI, MspI, and RsaI provided a tRF with the same length (score, 1.0) both for this bacterium and for *E. coli*. Two restriction enzymes had no site in the segment analyzed and therefore could not be considered. Despite the fact that the score was 0 with the two enzymes Hinfl and TaqI, the matches due to eight positive-testing enzymes were above the threshold level of two out of three. Thus, *S. gelidimarina* could not be excluded by the method employed. However, the phylogenetic tree clearly indicates that *S. gelidimarina* is at best distantly related to *E. coli* (85.1% homology between the two organisms). A similar calculation resulted in the selection of the other two giving false positives: *P. prydzensis* and *I. baltica*. In sum, however, the method employed is reliable in more than 95% of all cases, but care has to be taken to discern the few false positives. Thus, this approach allows retrieval not only of the composition of bacteria down to the genus level but also of most of the bacteria related to a specific organism (e.g., *E. coli*) from an environmental sample.

**FIG. 5.** The occurrence of the 16S rRNA gene tRF polygons related to those from *Azospirillum brasilense*, *Escherichia coli*, or *Rhizobium leguminosarum*. The TReFID program provided 192 polygons related to those of the three organisms and deposited in the TReFID databank (8 for *A. brasilense*, 111 for *E. coli*, and 64 for *R. leguminosarum*). Their sequences were used to construct the phylogenetic tree by use of the neighbor-joining method. The tree shows that only six artifacts were obtained (see text). The average sequence homologies between the three clusters of organisms (*A. brasilense*, *E. coli*, and *R. leguminosarum*) were 70.2, 80.8, and 71.1%, respectively. The numbers in the figure refer to the following: 1, *Roseomonas fauriae* (AF533354); 2, *Roseomonas* genomospecies (AY150050); 3, *Azospirillum* sp. (AB049110); 4, *Devosia riboflavina* (AF501346); 5, *Agrobacterium tumefaciens* (AF508094); 6, *Agrobacterium tumefaciens* (AF406666); 7, *Sinorhizobium kummerowiae* (AF364067); 8, *Sinorhizobium meliloti* (AF533685); 9, *Rhizobium tropici* (U89832); 10, *Mesorhizobium plurifarium* (AF516882); 11, *Ochrobactrum* sp. (AF452128); 12, *Ochrobactrum anthropii* (AF501340); 13, *Idiomarina baltica* (AJ440214); 14, *Xenorhabdus nematophila* (AF522294); 15, *Pectobacterium carotovorum* (AF373189); 16, *Serratia quinivorans* (AJ270050); 17, *Serratia* sp. (AF511524); 18, *Serratia odorifera* (AF286870); 19, *Pectobacterium carotovorum* (AF373184); 20, *Erwinia amylovora* (AF141892); 21, *Escherichia albertii* (AF508775); 22, *Escherichia coli* (AY319394); 23, *Ochrobactrum proteus* (AY077753); 24, "*Escherichia senegalensis*" (AY217654); 25, "Dickeya dadantii" (AF520707); 26, *Shewanella gelidimarina* (AF530149); 27, *Pseudoalteromonas* sp. (AB057888); 28, *Pseudoalteromonas prydzensis* (U85855).
Alphaproteobacteria represented about 60% of the total and thus the largest portion regarded as unclassified. These and other unclassified bacteria represented about 60% of the total and thus the largest portion in this soil (Table 4). The next-largest part consisted of α-proteobacteria. However, other major groups of soil bacteria that included acidobacteria and actinobacteria also occurred. Others (about 10%) consisted of spirochetes, firmicutes, bacteroidetes, fusobacteria, fibrobacteres, planctomycetes, δ-proteobacteria, and cyanobacteria; these latter were included in the list of the 16S rRNA gene sequences retrieved by using the TReFID program. However, sequences related to members of the Acidobacteria, Actinobacteria, and Pseudomonadales (including “sp.”) were also detected. The file obtained for bacteria with nosZ was small, but the highest percentage of these again belonged to the unclassified organisms. The actinobacteria and acidobacteria retrieved from the 16S rRNA gene sequences were not present in the list of bacteria with nifH or nosZ. Only a few bacteria, Bradyrhizobium japonicum, Mesorhizobium ciceri, Methyllosinus trichosporium, and Rhodovulum strictum, were found in the lists for both the 16S rRNA gene and nifH. Bacteria included in the list for the 16S rRNA gene and nosZ were Paracoccus denitrificans, P. pantotrophus, and Pseudomonas fluorescens. One bacterium, Pseudomonas stutzeri, possessed both nifH and nosZ but was not included in the list of the 16S rRNA gene.

<table>
<thead>
<tr>
<th>Category</th>
<th>16S rRNA</th>
<th>nifH</th>
<th>nosZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of matches</td>
<td>2,390</td>
<td>224</td>
<td>23</td>
</tr>
<tr>
<td>No. of different genera (classified in GenBank)</td>
<td>123</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>No. of different species (classified in GenBank)</td>
<td>227</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>No. of organisms not classified to genus level in GenBank</td>
<td>1,440</td>
<td>194</td>
<td>15</td>
</tr>
<tr>
<td>% Unclassified</td>
<td>60</td>
<td>87</td>
<td>65</td>
</tr>
</tbody>
</table>

TABLE 4. Taxa retrieved from the Dünnwald soil by using the TReFID databank

<table>
<thead>
<tr>
<th>Taxon</th>
<th>No. of species or genera for indicated gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16S rRNA</td>
</tr>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td>Alphaproteobacteria (including “sp.”)</td>
<td>431</td>
</tr>
<tr>
<td>Rhizobiales</td>
<td>192</td>
</tr>
<tr>
<td>Rhodobacterales</td>
<td>96</td>
</tr>
<tr>
<td>Rhodospirillales</td>
<td>28</td>
</tr>
<tr>
<td>Betaproteobacteria</td>
<td>37</td>
</tr>
<tr>
<td>Burkholderiales</td>
<td>32</td>
</tr>
<tr>
<td>Gammaproteobacteria</td>
<td>113</td>
</tr>
<tr>
<td>Pseudomonadales</td>
<td>9</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>52</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>31</td>
</tr>
<tr>
<td>Acidobacteria</td>
<td>3</td>
</tr>
</tbody>
</table>
terize N_2-fixing and denitrifying bacterial profiles in an environmental sample by a tRF analysis. The nifH gene coding for nitrogenase reductase is highly conserved among N_2-fixing bacteria (22), and more sequences are available for this gene than for the other two structural (nifDK) genes or for any other nif gene. The 822 polygons deposited in the TReFID databank are sufficient to allow characterization of the composition of a population of N_2-fixing bacteria in environmental samples, as shown for the Dünnwald soil, but the resolution will improve as more nifH sequences become available. Clearly, the situation is presently in infancy for nosZ or any other gene coding for a gene of denitrification. Thus, the present communication merely indicates that the technique can theoretically also be employed for denitrification.

The TReFID program presented here and the phylogenetic assignment tool recently published (6) should complement each other in characterizations of members of a bacterial community by their 16S rRNA gene tRF profiles. The two methods differ in several respects. The approach by Kent et al. (6) employs a hierarchical algorithm which allows the identification of organisms by analysis of the tRFs provided by the restriction enzymes in a consecutive order (see Fig. 1 of reference 6). Their tRF data have been automatically generated by MiCA (http://mica.ibest.uidaho.edu). The TReFID program utilizes polygons of tRFs of defined bacterial sequences obtained from databanks or a clone library. The TReFID program is presently more comprehensive for analysis of 16S rRNA gene sequences. It contains many sequences from bacteria not yet cultured. The program by Kent et al. (6) utilizes the 8F primer, whereas TReFID employs the 63F region, which is strongly conserved among the 16S rRNA gene sequences deposited in the databanks (13). In addition, many sequences of GenBank contain the 63F but not the 8F region. Only the use of the 63F region as a primer binding site enabled us to develop the more than 17,000 polygons used here for the 16S rRNA gene. The program described in reference 6 could not easily install such a detailed phylogram as that shown in Fig. 5 (C. Rösch, unpublished).

Previous molecular analyses of soil or marine samples suggested a relatively low species richness with respect to nifH but not the nosZ or the 16S rRNA gene (15, 16, 17, 21). The present study with DNA from the Dünnwald soil indicated by the use of the TReFID program that most nifH sequences retrieved were related to those of noncultured bacteria whose sequences have been deposited in the GenBank. It is presently not clear whether this result represents a special feature of the Dünnwald soil or whether the tRFPL analysis more readily accesses noncultured bacteria than other molecular approaches tried so far. To resolve this issue, a more detailed analysis of the bacterial population and its dynamics after N fertilization is presently under way with soil samples from Dünnwald and also from other locations. This is now possible, since the tRFPL TReFID program allows rapid analysis of the bacterial composition of an environmental sample and avoids time-consuming and expensive cloning and sequencing.

ACKNOWLEDGMENTS

We are indebted to M. G. Yates and G. B. Lewes, formerly at the Unit of Nitrogen Fixation of the ARC in Brighton, United Kingdom, for helpful comments and for improving the English of the manuscript. The excellent technical assistance of Emmanuelle Mounier, Stefanie Backhausen, Mirela Stecki, and Karin Otto is gratefully acknowledged. This study was kindly supported by grants from the GEW Stiftung (Cologne, Germany) and the Deutsche Bundesstiftung Umwelt (Osнabrack, Germany).

REFERENCES


Downloaded from http://aem.asm.org/ on March 28, 2021 by guest