APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2008, p. 6452–6456 Vol. 74, No. 20
0099-2240/08/$08.00+0 doi:10.1128/AEM.01394-08

TRiFLe, a Program for In Silico Terminal Restriction Fragment Length Polymorphism Analysis with User-Defined Sequence Sets▽

Pilar Junier,[1,3]*† Thomas Junier,[2]† and Karl-Paul Witzel[3]

Environmental Microbiology Laboratory, Ecole Polytechnique Federale de Lausanne, CH-1015 Lausanne, Switzerland[1]; Computational Evolutionary Genomics Group, University of Geneva, CH-1211 Geneva, Switzerland[2]; and Max Planck Institute for Evolutionary Biology, 24306 Ploen, Germany[3]

Received 21 June 2008/Accepted 19 August 2008

We describe TRiFLe, a freely accessible computer program that generates theoretical terminal restriction fragments (T-RFs) from any user-supplied sequence set tailored to a particular group of organisms, sequences from clone libraries, or sequences from specific genes. The program allows a rapid identification of the most polymorphic enzymes, creates a collection of T-RFs for the data set, and can potentially identify specific T-RFs in T-RF length polymorphism (T-RFLP) patterns by comparing theoretical and experimental results. TRiFLE was used for analyzing T-RFLP data generated for the *amoA* and *pmoA* genes. The peaks identified in the T-RFLP patterns show an overlap of ammonia- and methane-oxidizing bacteria in the metalimnion of a subtropical lake.

Terminal restriction fragment length polymorphism (T-RFLP) is a widely used molecular technique for studying microbial community composition and diversity in environmental (1, 3, 8) and clinical (2, 11) samples. For T-RFLP, PCR products (amplicons) are obtained using primers labeled with a fluorescent dye. Amplicons are digested with restriction enzymes, and the fragments generated are separated by high-resolution electrophoresis (e.g., in a DNA sequencer). The resulting fingerprint of the microbial community is the set of the lengths of all labeled terminal restriction fragments (T-RFs). T-RFLP analysis has been successfully applied for different targets, including 16S rRNA genes and genes of enzymes involved in specific metabolic processes, such as nitrogen fixation, nitrification, denitrification, or mercury resistance (7).

Specialized software can support the design and interpretation of T-RFLP experiments at two levels: (i) digestions of reference sequences can be simulated in silico in order to find appropriate enzymes for experimental analysis, and (ii) experimental T-RFLP patterns can be associated to predicted T-RFs from sets of reference sequences in order to identify possible species in the sample. Programs available on the web, such as MICA (microbial community analysis), TAP T-RFLP from the Ribosomal Database Project, or TReFID (6, 9, 12), can be used to perform in silico digestion of 16S rRNA genes. More recently, a similar module was integrated in the phylogenetic software program ARB (10). Although programs such as ARB can handle user-defined sets of sequences from genes other than 16S rRNA genes, this requires additional steps, such as the integration and alignment of the sequences, before the simulation can be performed. To our knowledge, none of the programs available so far has been specifically designed to simulate and create T-RF data sets using arbitrary sets of DNA sequences prepared from specific targets (e.g., genes involved in any metabolic pathways) or from unpublished sequences.

An increasingly popular trend in T-RFLP analysis consists of the identification of species in the samples by associating T-RFs from experimental runs with predicted T-RFs from a set of existing sequences. However, since related organisms commonly produce T-RFs of the same length, this association can be ambiguous, requiring digestion with several enzymes to increase the confidence on the assignment (5). Therefore, automation in the comparison of more-complex sets of data can contribute to the analysis and interpretation of T-RFLP data.

In this work we present the software program TRiFLe, which generates theoretical T-RFs from arbitrary sets of sequences by simulating PCR amplification and digestion with restriction enzymes. The main advantage of TRiFLe is thus that the simulation can be tailored to any desired groups of organisms, sequences from clone libraries, or specific genes. The results of the simulation can be used to design T-RFLP experiments or to compare theoretical and experimental T-RFs. The identification function included in TRiFLe allows the comparison of experimental results from several independent digestions with theoretical T-RFs from a data set of sequences. The program was validated by analyzing the diversity of ammonia- and methane-oxidizing bacterial communities in the metalimnion of Lake Kinneret (Israel) using PCR amplification, T-RFLP, and cloning of the genes *amoA* and *pmoA*.

Description of TRiFLe. TriFLe is a computer program written in Java and distributed as a Java Web Start application. This technology allows users to download and run the software by simply clicking on a link in a Web page and automatically handles updates. TRiFLe is available free of charge from the website at http://cegg.unige.ch/trifle/trifle.jnlp for the most common operating systems (Windows, Linux, and Mac OS). Its source code is distributed as open source software and includes a tutorial with examples for the application (http://cegg.unige.ch/trifle_docs).

Two different functionalities are implemented in the program. In the simulation function, the aim is to predict T-RFs

* Corresponding author. Mailing address: EPFL ENAC ISTE EML, CE 1 644 (Centre Est), Station 6, CH-1015 Lausanne, Switzerland. Phone: 41 21 693 63 96. Fax: 41 21 693 62 05. E-mail: pilar.junier@epfl.ch.

† These authors contributed equally to the manuscript.

▽ Published ahead of print on 29 August 2008.

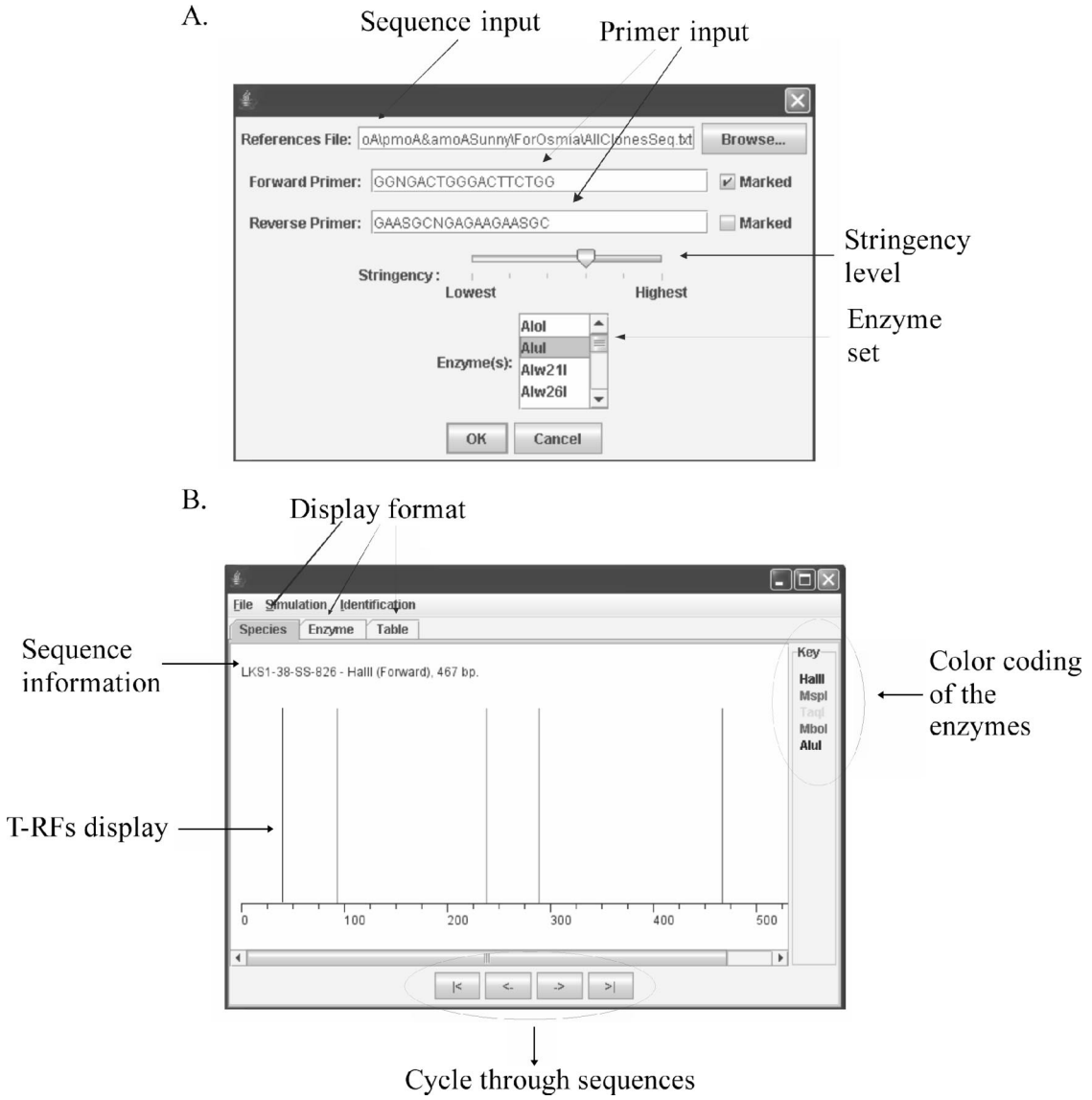

FIG. 1. Input interface of TRiFLe for simulating T-RFs (A) and graphical display of the results from the simulation (B). The different parts of the displays are indicated.

from sequences, primers, and enzymes given by the user. In the identification function, the program compares results from T-RFLP experiments with a data set of T-RFs from a set of reference sequences and computes a score to predict the community composition of the sample.

The input for the simulation of T-RFs (Fig. 1) consists of the following: (i) a FastA file containing the data set sequences, (ii) the primer sequences (these are just typed in a text field; IUPAC ambiguity codes can be used to specify degenerate primers), (iii) the labeled primers (forward, reverse, or both), and (iv) the set of restriction enzymes. The program constructs a probabilistic model (weight matrix) from each primer and searches the reference sequences for matches of each model. The matrix is slid along the candidate sequence, and each position is scored according to the matrix. The score is expressed as a probability. If the probability is above a certain threshold, which can be set by the user (through a slider in the dialogs), the position is considered a match. Nucleotide mismatches in the candidate sequence will lower the probability score, so the threshold gives control over the number of allowed mismatches.

Only sequences with matches of both primers ("amplicons") under the stringency conditions selected by the user are retained and digested in silico with the specified enzymes. For each digested amplicon, the program generates a graphical representation of all the predicted T-RFs (Fig. 1B). The results can also be displayed in other ways: as the set of all T-RFs generated by a particular enzyme or as a table containing the T-RFs from all the sequences in the data set. The results can be saved and loaded again and can be exported in TAB-separated format.

For the identification function, experimental profiles are compared with theoretical T-RF profiles generated from a set of sequences. The input data are as follows: (i) a FastA file

TABLE 1. Predicted and measured T-RFs of *nifH* sequences in five diazotrophic strains[a]

Strain or species	Length of T-RF (bp) with indicated restriction enzyme					
	Predicted			Measured		
	HaeIII	MspI	AluI	HaeIII	MspI	AluI
Frankia alni	46	236	150	46 ± 0.1	ND	149 ± 6.5
Rhizobium sp. strain NGR234	97	108	458	92 ± 0.3	102 ± 0.3	450 ± 1.6
Mesorhizobium loti	242	236	458	ND	232 ± 0.5	ND
Anabaena sp. strain PCC7120	458	206	46	436 ± 0.1	202 ± 0.3	39 ± 0.9
Bradyrhizobium japonicum	242	164	186	238 ± 0.4	161 ± 0.3	184 ± 0.9

[a] The experimental T-RFs were obtained by restriction with three different restriction enzymes with only the forward primer labeled. The experimental results are the average from five independent measurements. The standard deviation is indicated. ND, not determined.

containing the reference sequences, (ii) the primers, and (iii) a set of files from analyzed data of a T-RFLP experiment (run file), each containing experimentally measured T-RF lengths obtained with one enzyme using one labeled primer. For the run file, the program accepts any TAB-delimited table format and the user may define which of the columns correspond to the experimental fragment length, allowing run files with different formats to be analyzed. Considering that experimental

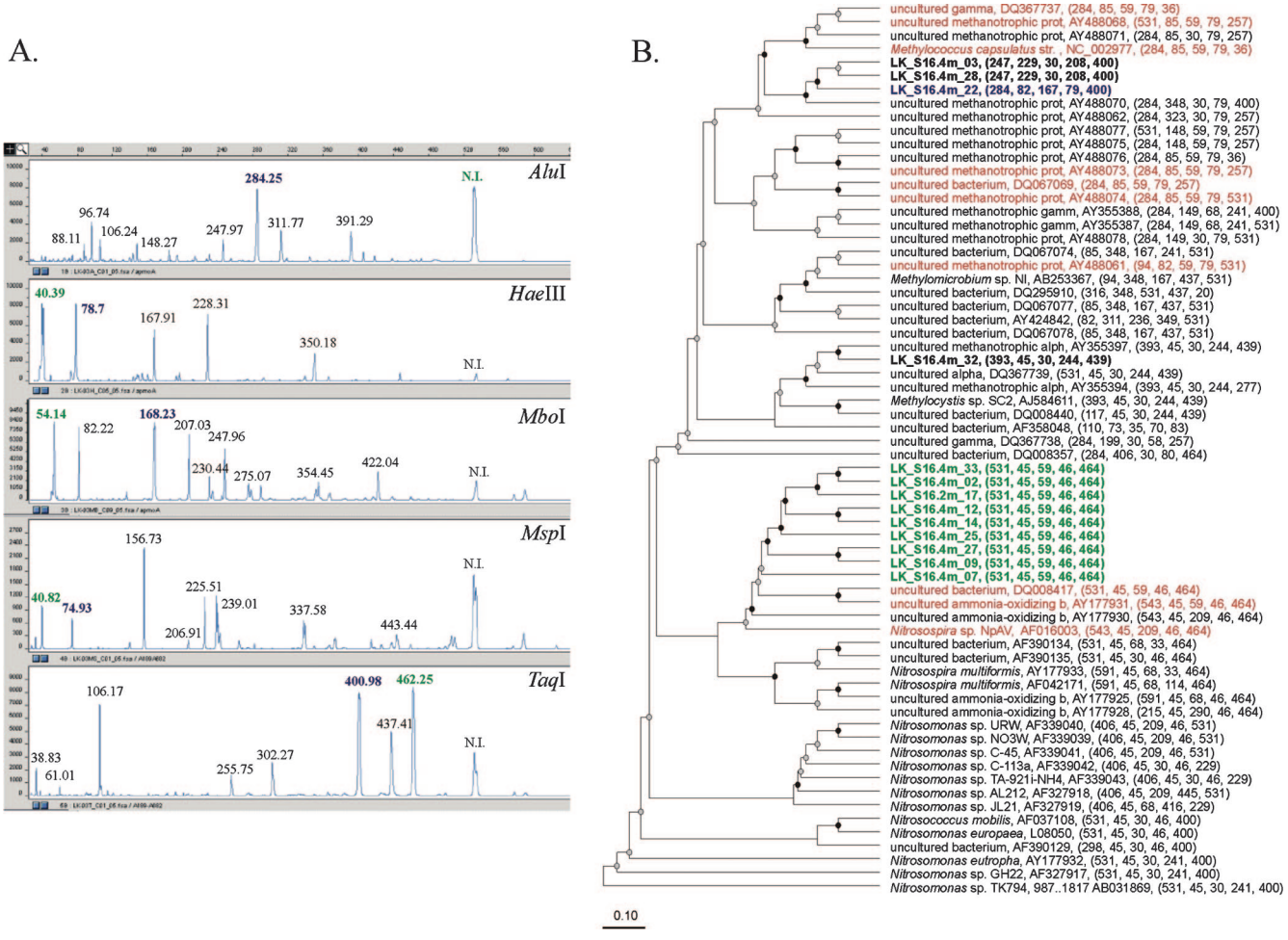

FIG. 2. Validation of the identification function of TRiFLe using experimental results from a T-RFLP experiment using a water sample at the metalimnetic layer of Lake Kinneret (Israel). (A) Electropherograms of the T-RFLP analysis of *pmoA* and *amoA* PCR products digested with HaeIII, MspI, MboI, AluI, and TaqI. T-RFs from a control set of clones that were identified by the program are shown in blue and green (colors indicate different phylogenetic groups). "N.I." indicates undigested peaks that were omitted for the calculation of the distance. (B) Phylogenetic tree of the reference data set of sequences used for the identification, including 13 clones from a library prepared from the environmental sample (bold). The simulated T-RF lengths calculated with TRiFLe for each of the enzymes are given in parentheses. The phylogenetic analysis was carried out using ARB (http://magnum.mpi-bremen.de/molecol/arb/); bootstraps values are indicated by black (100%) or gray (90 to 99%). The top 20 sequences identified by the program (see Table 2) are shown in color. Sequences in blue and green correspond to clones from the control set. Sequences in red correspond to reference sequences among the top 20. b, bacterium; prot, proteobacterium; alph or alpha, alphaproteobacterium; gamm or gamma, gammaproteobacterium; str., strain.

TABLE 2. Differences between observed *amoA* and *pmoA* T-RF sizes and those predicted using the identification function of TRiFLe[a]

Bacterium	GenBank sequence accession no.	Length or distance (bp) with indicated restriction enzyme															O (bp)
		AluI			HaeIII			MboI			MspI			TaqI			
		S T-RF	E T-RF	D	S T-RF	E T-RF	D	S T-RF	E T-RF	D	S T-RF	E T-RF	D	S T-RF	E T-RF	D	
Nitrosospira sp. strain NpAV	AF016003	543	395.17	N	45	43.70	1	209	209.02	0	46	44.12	2	464	469.80	N	1.0
Uncultured methanotrophic prot.	AY488068	531	395.17	N	85	81.68	3	59	57.34	2	79	77.94	1	257	257.29	0	1.5
Uncultured methanotrophic prot.	AY488073	284	285.89	2	85	81.68	3	59	57.34	2	79	77.94	1	257	257.29	0	1.6
Uncultured bacterium	DQ067069	284	285.89	2	85	81.68	3	59	57.34	2	79	77.94	1	257	257.29	0	1.6
LK_S16.4m_33		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_02		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_17		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_12		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_14		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_25		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_27		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_09		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
LK_S16.4m_07		**531**	**395.17**	**N**	**45**	**43.70**	**1**	**59**	**57.34**	**2**	**46**	**44.12**	**2**	**464**	**469.80**	**N**	**1.7**
Uncultured bacterium	DQ008417	531	395.17	N	45	43.70	1	59	57.34	2	46	44.12	2	464	469.80	N	1.7
Uncultured ammonia-oxidizing b	AY177931	543	395.17	N	45	43.70	1	59	57.34	2	46	44.12	2	464	469.80	N	1.7
Uncultured methanotrophic prot.	AY488074	284	285.89	2	85	81.68	3	59	57.34	2	79	77.94	1	531	469.80	N	2.0
LK_S16.4m_22		**284**	**285.89**	**2**	**82**	**81.68**	**0**	**167**	**170.24**	**3**	**79**	**77.94**	**1**	**400**	**405.24**	**5**	**2.2**
Uncultured methanotrophic prot.	AY488061	94	99.53	6	82	81.68	0	59	57.34	2	79	77.94	1	531	469.80	N	2.3
Methylococcus capsulatus strain	NC_002977	284	285.89	2	85	81.68	3	59	57.34	2	79	77.94	1	36	42.15	6	2.8
Uncultured gamma	DQ367737	284	285.89	2	85	81.68	3	59	57.34	2	79	77.94	1	36	42.15	6	2.8

[a] Results show the distance obtained for a data set of *amoA* and *pmoA* sequences compared with experimental T-RFLP data for a water sample at the metalimnetic layer of Lake Kinneret (Israel). Only the top 20 sequences are analyzed. Results for sequences of clones from a clone library prepared with the DNA of the sample used for the experimental T-RFLP are shown in bold (control set). S T-RF, simulated T-RF (length determined by TRiFLe); E T-RF, experimental T-RF used for comparison; D, distance between S T-RF and E T-RF lengths; O, overall distance considering all the enzymes simultaneously; N, undigested T-RF, omitted per the calculation of the distance; for bacterium abbreviations, see the legend for Fig. 2B.

lengths reported by a sequencer are known to be subject to errors (5, 7), the user can correct the experimental values using the correction formula of Kaplan and Kitts (5). Although this experimental correction was calculated for T-RFLP analysis using an ABI 310 genetic analyzer, it is so far the only experimental correction existing, and simulations with our data sets have shown good results when T-RFLP data from other systems have been corrected (data not shown).

For the identification of the T-RFs in the experimental samples, the program displays those T-RFs that were compared (simulated and experimental), as well as the distance (expressed in nucleotides). Additionally, considering that the experimental lengths of amplicons that do not contain an enzyme cut (unrestricted amplicons) are usually more biased (5), TRiFLe includes an option for setting a range of the fragments to be included in the calculation of the distance. Since different species may produce the same T-RF length with a particular enzyme and it is not possible to accurately quantify the contribution of each of them to the peak, a particular peak can be used in more than one identification. Therefore, having a larger set of enzymes can be expected to yield better identifications, since the overall distance is calculated from the combination of all the enzymes used.

Experimental validation of TRiFLe using known bacterial DNA. To test the program, results given by TRiFLe were compared with experimental results from T-RFLP of the nitrogenase iron protein gene (*nifH*) from the diazotrophic bacterial strains *Anabaena* sp. strain PCC7210, *Frankia* sp., *Bradyrhizobium japonicum* USDA110, *Rhizobium* sp. strain NGR234, and *Mesorhizobium loti* MAFF303099. The *nifH* gene was amplified using the primers nifHF and nifHR (13) from genomic DNA. For the amplification, the primer nifHF was labeled with 5-carboxyfluorescein. The amplicons were digested overnight at 37°C with 5 U of the restriction enzymes HaeIII and MspI (New England Biolabs) and afterwards separated on an ABI 3100 automatic sequencer. The predicted and observed T-RFs differed generally by 1% of the size of the fragment (Table 1). However, this difference was greater for fragments smaller than 50 bp or larger than 450 bp. The deviation between the predicted and observed T-RFs was in agreement with previous experimental determinations for 16S rRNA genes and *mrcA* (7).

Experimental validation using unknown microbial communities. TRiFLe's identification function was assayed using data generated from an environmental sample. A fragment of the genes coding for the alpha subunit of the particulate methane monooxygenase (*pmoA*) and ammonia monooxygenase (*amoA*) was amplified with the primer combination A189 and A682 (4) from DNA extracted from a water sample at the metalimnetic layer of Lake Kinneret (Israel), collected at station A (maximum depth, 42 m), which represents the pelagic area of the lake. The sample was filtered on 0.2-μm-pore-size filters (Supor-200; PALL Life Sciences) and stored at −18°C until DNA was extracted using the UltraClean soil

DNA kit (MoBio), following the manufacturer's guidelines. For T-RFLP, the primer A189 was labeled with 5-carboxyfluorescein. Three independent PCRs were pooled, gel purified, quantified, and digested overnight at 37 or 65°C with 5 U of HaeIII, MspI, MboI, AluI, and TaqI (New England Biolabs). Digest fragments were separated on an ABI 3100 automated sequencer. Fragment sizes were estimated by comparison with the ROX-500 standard (Applied Biosystems). The runs were analyzed using the GeneScan 3.1 software program (Applied Biosystems), and the resulting T-RF sets (Fig. 2A) were used for identification with TRiFLe.

Thirteen sequences from a clone library from the same environmental sample were included in the reference set to serve as a control set. The references also included *pmoA* and *amoA* sequences from cultured and uncultured methane- and ammonia-oxidizing bacteria reported in GenBank (Fig. 2B). The expected result was for TRiFLe to report the control set at or near the top of the identification list. As expected, TRiFLe reported most clones (10 out of 13) within the top 20 identified species (Table 2, bold entries). The T-RFs from those clones corresponded to the two most prominent peaks in the electropherograms (Fig. 2A). Several uncultured methanotrophic bacteria, *Methylocystis* sp. strain SC2, *Nitrosospira* sp. strain 39-19, and an uncultured ammonia-oxidizing bacterium were obtained in the top-20 list (Table 2). Those sequences were phylogenetically related to the clones in the control set (Fig. 2B). These results suggest that ammonia- and methane-oxidizing bacteria are colocalized in the layer of the metalimnion.

This research was supported by G.I.F. (German-Israel Foundation) grant no. I-711-83.8/2001 and BSF (Binational Science Foundation) grant no. 2002-206, and samples were taken during the German Israeli Minerva School in October 2004. We thank the Max Planck Society for financial support of P. Junier during this study.

We thank personnel of the Yigal Allon Kinneret Limnological Laboratory, Israel Oceanographic and Limnological Research, for their assistance during the sampling. We thank Ok-Sun Kim for testing the program and Ilonka Jäger, Tobias Lenz, Marco Pagnini, Dario Diviani, and Carlo Rivolta for their valuable comments.

REFERENCES

1. **Bruce, K. D.** 1997. Analysis of *mer* gene subclasses within bacterial communities in soils and sediments resolved by fluorescent-PCR-restriction fragment length polymorphism profiling. Appl. Environ. Microbiol. **63:**4914–4919.
2. **Christensen, J. E., J. A. Stencil, and K. D. Reed.** 2003. Rapid identification of bacteria from positive blood cultures by terminal restriction fragment length polymorphism profile analysis of the 16S rRNA gene. J. Clin. Microbiol. **41:**3790–3800.
3. **Clement, B. G., L. E. Kehl, K. L. DeBord, and C. L. Kitts.** 1998. Terminal restriction fragment patterns (TRFPs), a rapid, PCR-based method for the comparison of complex bacterial communities. J. Microbiol. Methods **31:** 135–142.
4. **Holmes, A. J., A. Costello, M. E. Lidstrom, and J. C. Murrell.** 1995. Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. FEMS Microbiol. Lett. **132:**203–208.
5. **Kaplan, C. W., and C. L. Kitts.** 2003. Variation between observed and true terminal restriction fragment length is dependent on true TRF length and purine content. J. Microbiol. Methods **54**:121–125.
6. **Kent, A. D., D. J. Smith, B. J. Benson, and E. W. Triplett.** 2003. Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities. Appl. Environ. Microbiol. **69:**6768–6776.
7. **Kitts, C. L.** 2001. Terminal restriction fragment patterns: a tool for comparing microbial communities and assessing community dynamics. Curr. Issues Intest. Microbiol. **2:**17–25.
8. **Liu, W. T., T. L. Marsh, H. Cheng, and L. J. Forney.** 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl. Environ. Microbiol. **63:**4516–4522.
9. **Marsh, T. L., P. Saxman, J. Cole, and J. Tiedje.** 2000. Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. Appl. Environ. Microbiol. **66:**3616–3620.
10. **Ricke, P., S. Kolb, and G. Braker.** 2005. Application of a newly developed ARB software-integrated tool for in silico terminal restriction fragment length polymorphism analysis reveals the dominance of a novel *pmoA* cluster in a forest soil. Appl. Environ. Microbiol. **71:**1671–1673.
11. **Rogers, G. B., C. A. Hart, J. R. Mason, M. Hughes, M. J. Walshaw, and K. D. Bruce.** 2003. Bacterial diversity in cases of lung infection in cystic fibrosis patients: 16S ribosomal DNA (rDNA) length heterogeneity PCR and 16S rDNA terminal restriction fragment length polymorphism profiling. J. Clin. Microbiol. **41:**3548–3558.
12. **Rösch, C., and H. Bothe.** 2005. Improved assessment of denitrifying, N_2-fixing, and total-community bacteria by terminal restriction fragment length polymorphism analysis using multiple restriction enzymes. Appl. Environ. Microbiol. **71:**2026–2035.
13. **Rösch, C., A. Mergel, and H. Bothe.** 2002. Biodiversity of denitrifying and dinitrogen-fixing bacteria in an acid forest soil. Appl. Environ. Microbiol. **68:**3818–3829.