APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2003, p. 7298–7309 Vol. 69, No. 12
0099-2240/03/$08.00+0 DOI: 10.1128/AEM.69.12.7298–7309.2003

Metagenome Survey of Biofilms in Drinking-Water Networks

C. Schmeisser,[1] C. Stöckigt,[1] C. Raasch,[2] J. Wingender,[3] K. N. Timmis,[4] D. F. Wenderoth,[4] H.-C. Flemming,[3] H. Liesegang,[2] R. A. Schmitz,[1] K.-E. Jaeger,[5] and W. R. Streit[1]*

Institut für Mikrobiologie und Genetik, Universität Göttingen,[1] and Laboratorium für Genomanalyse der Universität Göttingen,[2] 37077 Göttingen, Institut für Grenzflächenbiotechnologie, Universität Duisburg-Essen, 47057 Duisburg,[3] Institut für Molekulare Enzymtechnologie, Heinrich Heine-Universität Düsseldorf, Forschungszentrum Jülich, 52425 Jülich,[5] and Gesellschaft für Biotechnologische Forschung, 38124 Braunschweig,[4] Germany

Received 23 May 2003/Accepted 4 September 2003

Most naturally occurring biofilms contain a vast majority of microorganisms which have not yet been cultured, and therefore we have little information on the genetic information content of these communities. Therefore, we initiated work to characterize the complex metagenome of model drinking water biofilms grown on rubber-coated valves by employing three different strategies. First, a sequence analysis of 650 16S rRNA clones indicated a high diversity within the biofilm communities, with the majority of the microbes being closely related to the *Proteobacteria*. Only a small fraction of the 16S rRNA sequences were highly similar to rRNA sequences from *Actinobacteria*, low-G+C gram-positives and the *Cytophaga-Flavobacterium-Bacteroides* group. Our second strategy included a snapshot genome sequencing approach. Homology searches in public databases with 5,000 random sequence clones from a small insert library resulted in the identification of 2,200 putative protein-coding sequences, of which 1,026 could be classified into functional groups. Similarity analyses indicated that significant fractions of the genes and proteins identified were highly similar to known proteins observed in the genera *Rhizobium*, *Pseudomonas*, and *Escherichia*. Finally, we report 144 kb of DNA sequence information from four selected cosmid clones, of which two formed a 75-kb overlapping contig. The majority of the proteins identified by whole-cosmid sequencing probably originated from microbes closely related to the alpha-, beta-, and gamma-*Proteobacteria*. The sequence information was used to set up a database containing the phylogenetic and genomic information on this model microbial community. Concerning the potential health risk of the microbial community studied, no DNA or protein sequences directly linked to pathogenic traits were identified.

Current estimates indicate that more than 99% of the microorganisms present in many natural environments are not readily culturable and therefore not accessible for biotechnology or basic research (1). In fact, most of the species in many environments have never been described, and this situation will not change until new culture technologies are developed (1). Additionally, many approaches currently used to explore the diversity and potential of microbial communities are biased because of the limitations of cultivation methods.

To overcome the difficulties and limitations associated with cultivation techniques, several DNA-based molecular methods have been developed. In general, methods based on 16S rRNA gene analysis provide extensive information about the taxa and species present in an environment. However, these data usually provide little information about the functional role of any of the different microbes within the community and the genetic information they contain.

Metagenomics is a new and rapidly developing field that tries to analyze the complex genomes of microbial niches. Although the term metagenome has been introduced only recently to describe the genomes of noncultivated microbes present within a soil microbial community (10), earlier studies used a similar approach. In one such study, the approach was employed for the isolation of cellulases from a thermophilic environment (11), and in a second study the approach was used for the phylogenetic characterization of marine picoplankton (27).

Since then, an increasing number of publications have applied similar techniques to study the metagenomes of diverse microbial communities. The microbial niches addressed within these studies included the characterization of a wide range of different microbial communities ranging from soil and rather extreme environments to laboratory enrichments (2–5, 7, 19, 22, 24, 25, 32). The goal of these studies was to increase our understanding of ecological and molecular processes in the microbial communities, and several of these studies also aimed at an increased understanding of the genome information of individual microbes within the complex communities. In addition, the approach has been used to identify a number of novel biocatalysts and other interesting biomolecules from noncultivated microbes (8, 9, 11–13, 16, 17, 32). Altogether, these studies have led to an increased knowledge of the genetic structure of the microbial communities studied. Despite the number of metagenome studies, the amount of DNA information generated for individual niches is still very limited if one takes into account that the DNA information of several thousand different microbial genomes may be stored within a single microbial habitat (31). Thus, conclusions on the functional role of the microbes and sequences identified within these highly diverse bacterial communities cannot easily be made.

Since it can be assumed that microbial biofilms commonly found in drinking water distribution systems typically consist of

* Corresponding author. Mailing address: Institut für Mikrobiologie und Genetik, Universität Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany. Phone: (49) 551-393775. Fax: (49) 551-393793. E-mail: wstreit@gwdg.de.

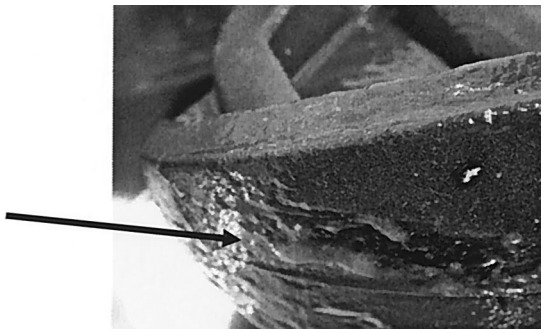


FIG. 1. Bacterial biofilm observed on the surface of a rubber-coated drinking water valve. The arrow indicates the bacterial biofilm. The drinking water valve was obtained from a drinking water pipe with an internal diameter of 15 cm. The valves are normally submerged in the drinking water.

fewer bacterial species than soil samples, they are ideal models to study metagenomes in combination with a phylogenetic analysis. The microbial communities that build drinking water biofilms have been characterized to some extent by 16S rRNA gene analyses. While these studies have mostly focused on the detection of bacterial species causing infectious diseases, such as *Legionella* and indicator organisms for fecal contamination, such as coliform bacteria (30), a number of more recent studies have led to the identification of novel nonpathogenic bacterial species (14, 15). Thus, the metagenomes of drinking water biofilms represent distinct and highly intriguing ecological niches, and their analysis is of significance to both the water suppliers and the consumers.

The aim of this study was to give insight into the metagenomes of drinking water biofilms grown on rubber-coated valves. For this purpose we characterized the phylogenetic structure of bacterial biofilms derived from rubber-coated drinking water valves by sequencing 16S rRNA clones. Additionally, we generated and analyzed about 2.0 Mb of DNA sequence information with a snapshot genome sequencing approach. With this sequence information, we analyzed the DNA sequence of four cosmid clones. This information has been used to set up a database to link the phylogenetic information with the genomic and functional information and to shed new light on the fine structure and evolution of the metagenomes of such complex microbial communities.

MATERIALS AND METHODS

Total DNA extraction. For the analysis, three biofilm samples were collected within the drinking water networks of a town in the northwestern part of Germany in the state of North Rhine-Westphalia, and the samples were all obtained from the surfaces of identical ethylene-propylene-diene monomer-coated valves. Prior to removal, the rubber-coated valves were submerged in nonchlorinated drinking water for 4 to 7 months. Samples were frozen at −70°C until processing. For library construction, three samples were collected from the surface of the rubber-coated valves (Fig. 1), and the samples were designated BioI, BioII, and BioIII. Total nucleic acids were extracted from the biofilms by standard protocols (8).

Cosmid and small insert libraries were constructed as previously published (8). After collection, bacteria were resuspended in TE-sucrose (20%, wt/vol) buffer and lysed in DNA extraction buffer (100 mM Tris HCl, 100 mM EDTA, 100 mM Na_2HPO_4, 1.5 M NaCl, 1% SDS) for several hours. RNA was degraded with RNase A (10 mg/ml). The resulting DNA extracts were incubated with protease and Sarcosyl (5%, wt/vol) in TE buffer overnight. Total genomic DNA was then repeatedly extracted with chloroform-phenol (1:1, vol/vol), washed once with chloroform, and dialyzed against 2 liters of TE buffer at 4°C overnight. Finally, an aliquot of the DNA was analyzed on a 0.8% agarose gel to ensure that the DNA was not degraded.

Cosmid libraries were prepared in pWE15 (Stratagene, La Jolla, Calif.) with standard protocols (8). DNA fragments (20 to 40 kb) obtained after partial *Sau*3A digestion were ligated into the *Bam*HI restriction sites of the cosmid vector. Phage packaging mixes were obtained from Stratagene (La Jolla, Calif.), and infection of *Escherichia coli* VCS257 was performed according to the manufacturer's protocol. For the construction of the snapshot libraries, DNA fragments with inserts of 3 to 7 kb were ligated into the sequencing vector pTZ19R (Amersham-Pharmacia, Essex, United Kingdom) and transformed into *E. coli*. For the construction of cosmid and small-insert libraries, the DNAs of the three samples were pooled. This was necessary because the amounts of DNA obtained from each individual sample were not sufficient to allow construction of the different samples. Therefore, the DNA of the three samples is considered a pool of biofilm genomes throughout this work, and the data summarize the possible microbes and genes occurring in these microbial niches.

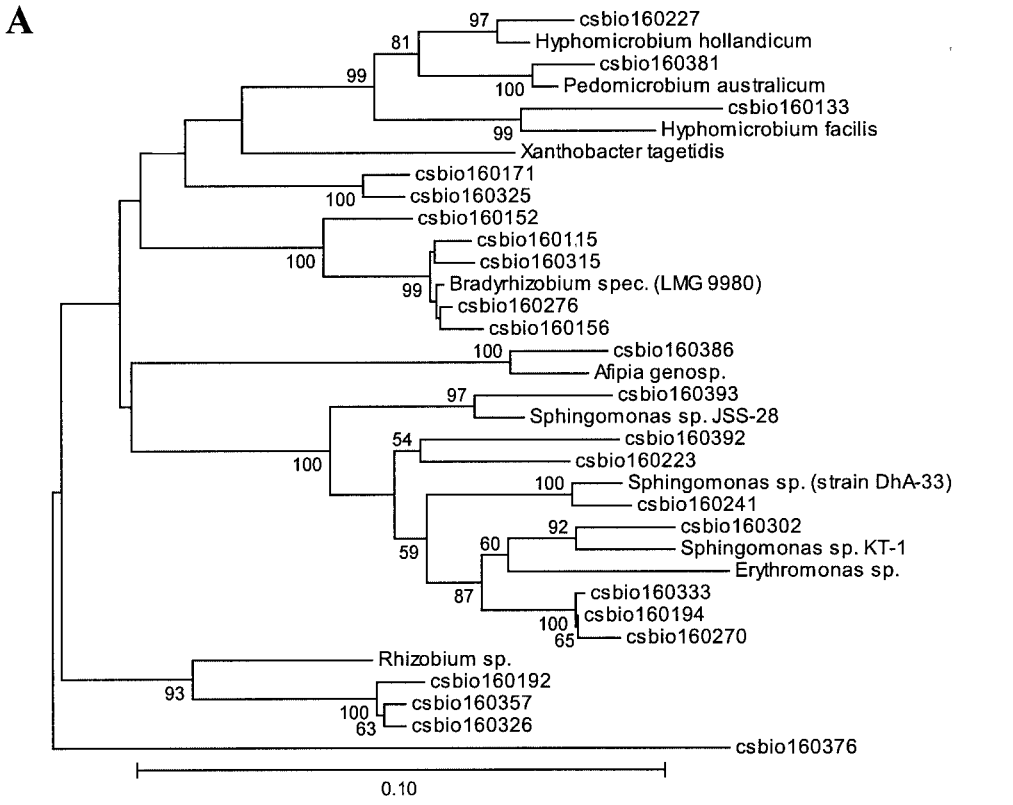
PCR and cloning of 16S rRNA sequences. Bacterial biofilm ribosomal DNAs (rRNAs) were amplified by PCR from DNA in reaction mixtures containing (as final concentrations) 1× PCR buffer (Perkin-Elmer), 2.5 mM $MgCl_2$, 200 μM each deoxynucleoside triphosphate, 300 nM each forward and reverse primer, and 0.25 U of *Taq* DNA polymerase (Perkin-Elmer) per ml. Reaction mixtures were incubated in a gradient thermal cycler (MJ Research, Boston, Mass.) at 96°C for 5 min for initial denaturation, followed by 25 to 35 cycles at 94°C for 30 s, 50°C for 45 s, and 72°C for 1.5 min, followed by a final extension period of 10 min at 72°C. For the clone library, rRNA genes were amplified with the universal reverse oligonucleotide primer 5′-CGGCCTCTACCTTGTTACGAC-3′ and the universal forward primer 5′-AGAGTTTGATCCTCACTGGCTCAG-3′. The resulting PCR products (of 1.5 kb) were cloned with a Topo TA cloning kit in accordance with the manufacturer's instructions (Invitrogen Corp., Karlsruhe, Germany). Plasmid DNAs containing inserts were sequenced with standard protocols for ABI 377 automated sequencing.

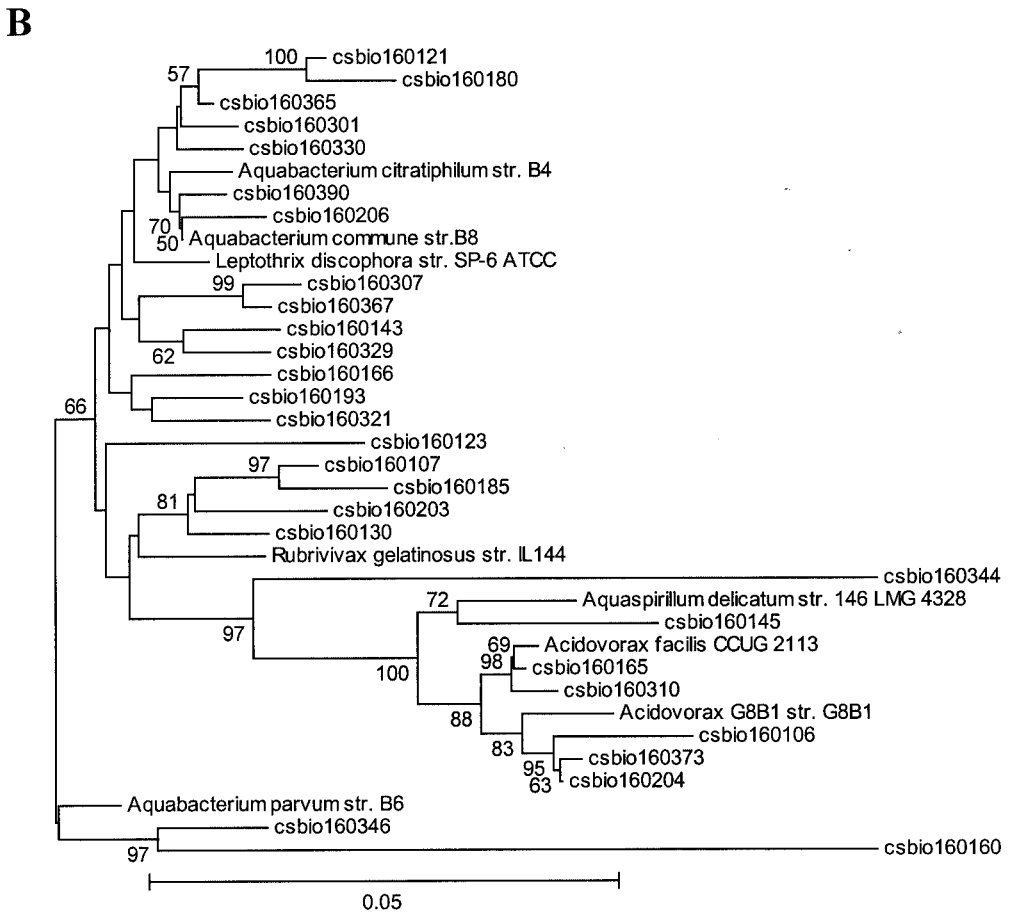
Assignment of cloned sequences to established phylogenetic divisions. The phylogenetic diversity was assessed with clone libraries of the 16S rRNA gene sequences of the different biofilm samples. The cloned 16S rRNA gene sequences were compared with reference sequences contained in the NCBI nucleotide sequence database with the FASTA program. For calculation of a phylogenetic tree, all ambiguous positions were excluded from similarity calculations. Sequences were screened for chimeras with the Check_Chimera program of the Ribosome Database Project and by manual alignments of secondary structure. As a final check for chimeras, each sequence was split into 5′ and 3′ fragments, which were analyzed separately by Blast searching of GenBank. Sequences for which either the 5′ or 3′ fragment had significantly different closest relatives were considered probable chimeras and were removed from the data set.

For calculation of the dendrogram shown in Fig. 2, cloned sequences were aligned with 16S rRNA gene sequences representative of the main bacterial divisions. Sequences were aligned with 16S rRNA sequences of other bacteria obtained from the Ribosomal Database Project (RPD-II) (18). Matrices of evolutionary distance were computed from the sequence alignment with the program DNADIST implemented in the software package Phylip (http://evolution.genetics.washington.edu/phylip.html) (version 3.5). For calculations of a phylogenetic tree from the distance matrices, the program applies the neighbor-joining method described by Saitou and Nei (23).

Single-strand conformation polymorphism analysis. The single-strand conformation polymorphism analyses of the biofilm communities were done following standard protocols (26, 29). After PCR amplification of the partial the 16S rRNA genes with primers COM1 (5′CAGCAGCCGCGGTAATAC3′, positions 519 to 536) and reverse primer COM2-Ph (5′CCGTCAATTCCTTTGAGTTT3′, positions 907 to 926) (29), the phosphorylated strand of the amplified PCR fragments was removed by λ exonuclease digestion. The fragment size of the amplified V4 and V5 regions of the 16S rRNA gene was 390 bp. To introduce specific secondary structures in the strands, samples were heat denatured, quickly chilled on ice, and then electrophoresed on nondenaturing gels; bands visualized by silver staining. For determining the band numbers, the gels were digitized to create TIF files. Analysis of the 16S rRNA fingerprints was performed with the software package GelCompare II (Applied Maths, Kortrijk, Belgium). The background was subtracted with rolling-circle correction (circle diameter, 30 points), and lanes were normalized. Only bands with an intensity of 2% or more of the total lane intensity were considered.

Nucleotide sequence data analysis. Automated DNA sequencing was performed with ABI377 and dye terminator chemistry following the manufacturer's instructions; when required, gaps in the DNA sequences were filled by PCR. The nucleotide sequences obtained for larger contigs or complete cosmids have been


A
csbio160227
Hyphomicrobium hollandicum
csbio160381
Pedomicrobium australicum
csbio160133
Hyphomicrobium facilis
Xanthobacter tagetidis
csbio160171
csbio160325
csbio160152
csbio160115
csbio160315
Bradyrhizobium spec. (LMG 9980)
csbio160276
csbio160156
csbio160386
Afipia genosp.
csbio160393
Sphingomonas sp. JSS-28
csbio160392
csbio160223
Sphingomonas sp. (strain DhA-33)
csbio160241
csbio160302
Sphingomonas sp. KT-1
Erythromonas sp.
csbio160333
csbio160194
csbio160270
Rhizobium sp.
csbio160192
csbio160357
csbio160326
csbio160376
0.10


B
csbio160121
csbio160180
csbio160365
csbio160301
csbio160330
Aquabacterium citratiphilum str. B4
csbio160390
csbio160206
Aquabacterium commune str.B8
Leptothrix discophora str. SP-6 ATCC
csbio160307
csbio160367
csbio160143
csbio160329
csbio160166
csbio160193
csbio160321
csbio160123
csbio160107
csbio160185
csbio160203
csbio160130
Rubrivivax gelatinosus str. IL144
csbio160344
Aquaspirillum delicatum str. 146 LMG 4328
csbio160145
Acidovorax facilis CCUG 2113
csbio160165
csbio160310
Acidovorax G8B1 str. G8B1
csbio160106
csbio160373
csbio160204
Aquabacterium parvum str. B6
csbio160346
csbio160160
0.05

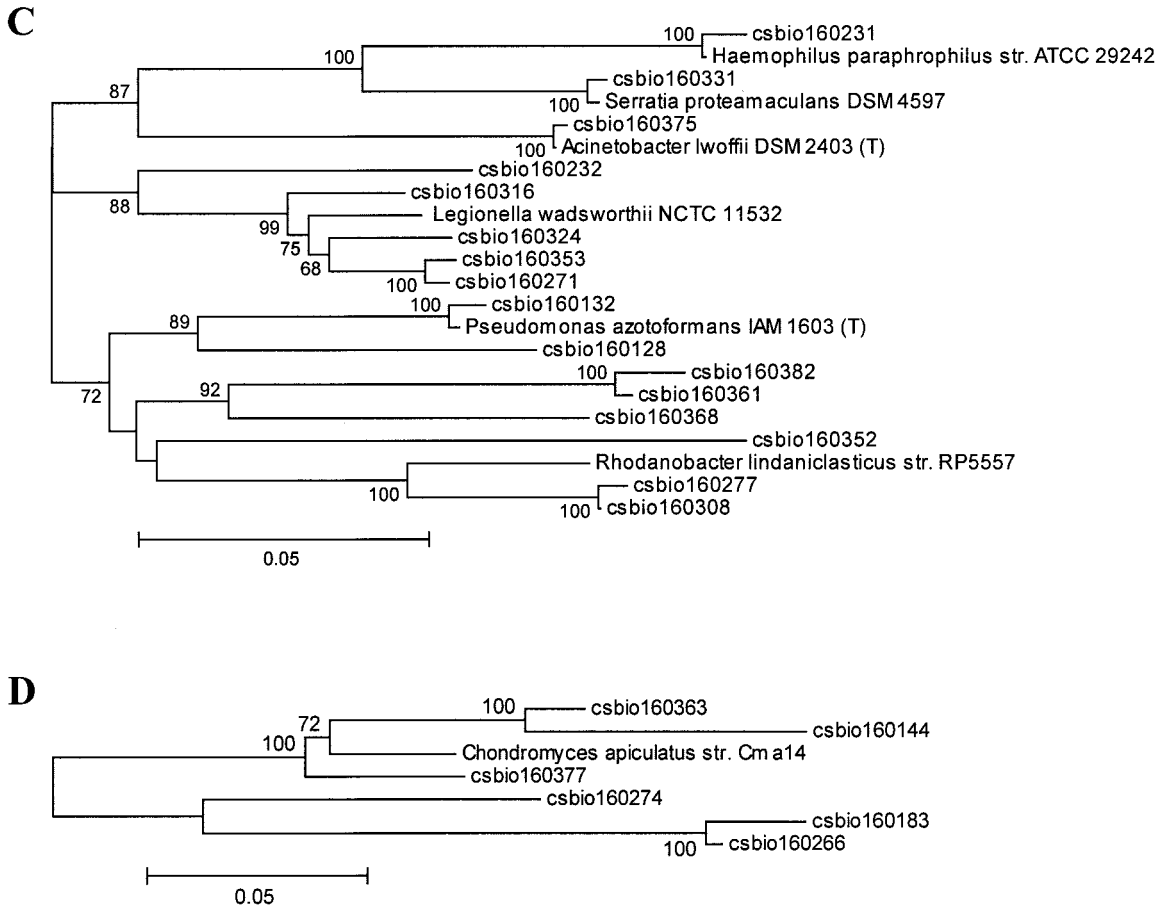


FIG. 2. Dendrogram of the 16S rRNA clones identified within the drinking water bacterial community DNA, showing the relationship to the closest known relatives. Only the proteobacterial lineages species are depicted. (A) Alpha-proteobacterial lineage; (B) beta-proteobacterial lineage; (C) gamma-protebacterial lineage; (D) delta-proteobacterial lineage. The phylogenetic trees were calculated with the software package MEGA version 2.1 (Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Ariz.) and verified with the Phylip software package from the Ribosomal Database Project (RDP) (18). Only high-quality sequences from the 16S rRNA gene clones were included in the calculations, and the hypervariable regions in the 16S rRNA molecule were excluded from the calculations. Numbers indicate data from a bootstrap analysis, and values below 50% are not indicated.

deposited at GenBank, and accession numbers are listed in Tables 4 to 6. The sequence data of the cloned 16S rRNA genes were deposited at GenBank, and all 81 accession numbers (AY187312 to AY187393) are available at www.gwdg.de/~biofilm/together with the corresponding sequences; the snapshot genome sequences are available at the same web pages together with the BlastX results. Also, the sequences of the completely sequenced cosmids are available together with the GenBank accession numbers and other useful information on this web site. The GC contents of the nucleic acid sequences from the cosmids and the snapshot library was calculated with the program Geecee from the free open-source software package for sequence analysis, Emboss (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) running on a local Linux server.

RESULTS

Phylogenetic analyses. The samples used for total nucleic acid extraction were taken from the surfaces of the rubber-coated drinking water valves and used for DNA extraction (Fig. 1). DNAs of three biofilms, which were grown on the rubber coated-surfaces were pooled. The phylogenetic diversity of the bacterial biofilm community was assessed with the cloned and pooled rRNA gene sequences of the pooled biofilm samples. For this purpose, 650 clones were analyzed, and this resulted in the identification of 81 different clones. These sequences are phylogenetically highly diverse and include numerous bacterial lineages (Fig. 2).

Interestingly, no single phylogenetic group of bacteria dominated the clone collection. Instead, common bacterial phylotypes that occurred in the sample included members of the alpha-, beta-, delta-, and gamma-*Proteobacteria*, the *Cytophaga-Flavobacterium-Bacteroides* group, the *Actinobacteria*, and the low G+C gram-positive group (Fig. 2A to D and Table 1). Altogether, the *Proteobacteria* constituted 86% of the clones identified and thus represented the largest fraction of microbes within the bacterial community. The *Actinobacteria*, the low G+C gram-positives, the *Cytophaga-Flavobacterium-Bacteroides* group, and the *Acidobacteria* constituted only minor fractions of the clones. Finally, a small number of sequences were highly similar to unclassified bacteria (Table 1). While several of the isolates were highly similar to previously described microbial species within drinking water bacterial

TABLE 1. Different phylogenetic groups and clones observed in the 16S rRNA clone library derived from a drinking water biofilm community DNA[a]

Bacterial division	No. of clones	% of total
Total	81	100
Proteobacteria	70	86
Alpha subdivision	23	28
Beta subdivision	29	36
Gamma subdivision	15	19
Delta subdivision	3	4
Other groups	11	14
Actinobacteria	1	1.2
Low G+C gram-positives	1	1.2
Cytophaga-Flavobacterium-Bacteroides	4	4.9
Acidobacteria	3	3.7
Unclassified bacteria	2	2.5

[a] Clones were significantly different when DNA similarities were lower than 97%. Data were generated by sequence analysis of 450 clones, resulting in the identification of 81 different clones.

communities, a novel observation was that a limited number of the clones identified were closely related to the microbes which belong to the genera *Rhizobium* and *Bradyrhizobium*.

Further tests were employed to verify the high phylogenetic diversity within the microbial communities studied. For this purpose, single-strand conformation polymorphism genetic profiles of the different drinking water biofilm microbial communities DNA were analyzed. Primers designed to amplify the bacterial 16S rRNA gene sequences, including the variable V4 and V5 regions, yielded complex single-strand conformation polymorphism patterns on polyacrylamide gels. In these tests, the observed profiles consisted of more than 35 different product bands for each of the samples tested (data not shown).

Random sequencing of 2,500 small insert clones containing biofilm DNA. Total genomic DNA of the drinking water biofilms was used to construct a small insert library with inserts ranging in size from 1 to 5 kb. Of the 5,000 random sequences obtained, 2,496 produced high-quality DNA sequences (Table 2); and 2,504 sequences (50.1%) were not included in further analyses because of poor sequence quality, short length of the reads, or vector contaminations. In this way, more than 2.0 Mb of high-quality nucleotide sequence were collected and analyzed. The G+C content of the high-quality sequences was 62%.

TABLE 2. Overview of snapshot genome sequence analysis of a small insert library of drinking water biofilm DNA[a]

Sequence type	No.	% of total
Sequences generated	5,000	100.0
High-quality sequences over 800 bp in length	2,496	49.9
Low-quality sequences (not included in further analysis)	2,504	50.1
Sequences with significant similarity (E value $< 10^{-4}$)	1,344	26.9
Sequences with weak similarity (E value $>10^{-4}$)	856	17.0
Sequences with no hit in database	296	6.0

[a] High-quality sequences refers to sequence more than 800 bp in length and a confidence value of >15. Sequences were analyzed by automated BlastX search at the NCBI nonredundant databases. An E value of 10^{-4} was arbitrarily chosen as the cutoff for similarity searches.

TABLE 3. Functional classes and possible ORFs identified in random biofilm genome sequences after automated BlastX searches[a]

Functional class	No. of complete or partial proteins identified	% of total
Regulatory function	112	8.4
Metabolism and catabolism	455	34.0
Cell processes and structure	132	9.0
Elements of external origin	22	1.7
DNA/RNA-modifying enzymes	117	8.8
Protection responses	28	2.1
Transport proteins	112	8.4
Hypothetical proteins	318	23.9
Miscellaneous	48	3.7
Total	1,344	100.0

[a] All BlastX results and corresponding sequences are publicly available together with other information at http//www.gwdg/~biofilm/.

To assign putative functions to the cloned DNA fragments, sequences were compared to the NCBI protein and nucleotide databases. BlastX analyses indicated that 1,344 of the 2,496 high-quality sequences matched known protein-coding ORFs (Table 2). Of the 1,344 putative protein-coding sequences, 318 (24%) were similar to hypothetical genes with no known function (Table 3). BlastX searches with 296 of the sequences did not return any significant similarities. To provide an overview of the genetic organization of the biofilm metagenome, 1,344 predicted protein-coding sequences, based on BlastX searches, were grouped into nine classes according to their putative function (Table 3). Also, all BlastX results are available at http://www.gwdg.de/~biofilm.de together with the corresponding sequences and other information on the metagenome analyzed.

Catabolic and metabolic abilities stored in the biofilm metagenome. A total of 455 (34%) sequences were found to encode putative proteins involved in catabolic or metabolic activities of the microbial biofilm community (Table 3). Of these, the majority encoded genes involved in classical pathways such as the tricarboxylic acid cycle, 2-keto-3-deoxy-6-phosphogluconate pathway, glycolysis, and the glyoxylate cycle. Interestingly, quite a large number of possible genes involved in lipid hydrolysis could be identified. Altogether, 21 partial genes coding for possible lipases were identified, suggesting that lipolytic activities are probably of importance for this biofilm community. Most of the putative lipases were highly similar to lipases known to be present in *Pseudomonas fluorescens*.

Furthermore, a number of genes were identified which encoded proteins involved in the degradation of aromatic compounds. These included mostly genes involved in the degradation of toluate and benzoate or related compounds. The partial proteins were highly similar to corresponding proteins from gram-positive and gram-negative microbes. Also, 14 possible ORFs were identified encoding genes involved in the degradation or modification of polysaccharides (i.e., starch and cellulose). Surprisingly, 21 putative protease genes were identified and 12 ORFs possibly involved in the catabolism of amino acids were found. Altogether, these findings suggest that the microbial community analyzed in this study is nutritionally

highly diverse and able to catabolize a wide range of different carbon and energy sources.

Other remarkable features included the identification of 28 (2.1%) sequences encoding genes that are involved in protection response, such as antibiotic resistance or metal detoxification. Eight clones carried possible tetracycline resistance genes, and seven clones were possibly involved in resistance to β-lactam antibiotics. Two ORFs were identified that might be linked to bacterial polyketide synthesis. Other features identified included possible ORFs involved in bacterial photosynthesis and light emission. Finally, it is noteworthy that none of the sequences of the snapshot analysis encoded proteins specifically related to pathogenic mechanisms. A complete list of all the possible ORFs identified and their possible functions is available at http://www.gwdg.de/~biofilm/overviewtable.htm.

Statistical and phylogenetic analysis of the BlastX hits. To further exploit the DNA snapshot sequences, we analyzed the distribution of BlastX hits over different bacterial groups. For this purpose, the results of 1,026 BlastX similarity searches were evaluated. The statistical analysis of the BlastX searches indicated that the major fraction (84%) of all proteins were highly similar to proteins derived from the *Proteobacteria* (Fig. 3A). Among these, most were highly similar to the group of the alpha- and gamma-*Proteobacteria* (74.3%). Among the proteins most similar to proteins originating from the alpha-*Proteobacteria*, the largest fraction were highly similar to rhizobial proteins (i.e., *Rhizobiales*) (Fig. 3B). Interestingly, within the *Rhizobiales* most deduced proteins were highly similar to *Sinorhizobium meliloti* and *Mesorhizobium loti* proteins (Fig. 3C). Also, a significant fraction of proteins (5.5%) were highly similar to proteins originating from microbes closely related to the typical freshwater microbe *Caulobacter crescentus*.

Within the group of the gamma-*Proteobacteria*, the majority of proteins were highly similar to proteins derived from bacteria closely related to the *Pseudomonadales* (14.7%) and *Enterobacteriales* (10.8%) (Fig. 3B). The possible ORFs identified within the *Pseudomonadales* were highly similar to proteins derived from *Pseudomonas aeruginosa* and *P. fluorescens.* Furthermore, 6.8% of all proteins analyzed appeared to be highly similar to proteins originating from microbes related to *Ralstonia solanacearum*. Finally, it is noteworthy that 7.2% of all proteins analyzed were highly similar to known proteins from the *Actinomycetales* (i.e., *Streptomyces* and *Mycobacterium*) (Fig. 3B and C). Altogether, the statistical analysis of the putative ORFs supports the idea that the biofilm community studied is highly diverse. The data also suggest that a significant number of the proteins possibly expressed in the bacterial community originates from microbes closely related to the *Rhizobiales*.

Sequence analysis of large insert clones. To further exploit the genomic information contents of drinking water biofilms, the complete DNA sequences of four cosmid clones were determined. Three of the sequenced cosmid clones were randomly selected from a library containing approximately 2,500 clones, and the sequenced clones were designated pbioW, pbioV, and pbioX. Cosmid clone pbioY was selected because it overlapped cosmid pbioX. In total, 144 kb of additional DNA sequence information was generated, and this resulted in the identification of 94 ORFs. The G+C content was highly similar for all the cosmids and ranged between 65 and 67%. The nucleotide sequences obtained for the cosmids have been deposited at GenBank, and the accession numbers are listed in Tables 4 to 6. All ORFs identified on the sequenced cosmids are summarized in Fig. 4.

The insert size of pbioV was 37.8 kb, and the cosmid encoded 22 ORFs; 13 ORFs encoded hypothetical proteins, and many of these were probably involved in cellular processes. Two genes were identified which were involved in the biosynthesis of panthothenate (*panB* and *panC*), and two ORFs possibly involved in amino acid biosynthesis (*csv020* and *aroC*) were identified (Table 4).

Cosmid clone pbioW encoded 22 ORFs in its 30.8-kb insert. Among these was a cluster of ORFs possibly involved in nitrogen regulatory circuits. Other possible genes encoded included a heme oxygenase and two proteins possibly involved in DNA modification. In addition, a number of hypothetical proteins were identified (Table 5).

DNA restriction analysis and sequencing indicated that cosmids pbioX and pbioY formed a 75-kb contig of biofilm DNA. Altogether, 51 ORFs were identified through the DNA sequence analysis. Among the possible genes identified were mostly genes involved in cellular processes. Also, one possible transposase (*csx031*) and several regulatory genes were identified. Additionally, we encountered at least five different ORFs with potential value for biotechnological application. ORFs *csx002* and *csx024* encoded putative novel lipases, and ORFs c*sx006*, *csx007*, and *csx008* encoded putative amylolytic enzymes. Finally, three ORFs encoding a possible drug resistance transporter were identified (*csx012* to *csx014*) (Table 6).

Of the 94 identified proteins, three were highly similar to proteins derived from delta-*Proteobacteria*, 14 were highly similar to proteins derived from the alpha-*Proteobacteria*, 34 were highly similar to the beta-proteobacterial proteins, and 30 were highly similar to proteins derived from gamma-proteobacterial species. Only 13 proteins were highly similar to known proteins from gram-positive microbes or other microbial species (Fig. 4). Altogether, the analysis of large insert clones also supports the concept that the studied biofilm is mainly constructed of microbes closely related to known species of the alpha-, beta-, and gamma-proteobacterial lineages.

In summary, all these data give a first insight into the complex metagenome of biofilms derived from rubber-coated valves used in drinking water networks.

DISCUSSION

The primary focus of the present paper was to provide high-resolution information on the genome information stored within the metagenome of drinking water biofilms grown on rubber-coated valves. This was achieved with three different strategies. Our first approach included an analysis of the 16S rRNA genes of the microbes present within the microbial community. The phylogenetic data indicated that the microbial community is constructed out of a significant number of different and mostly nonpathogenic proteobacterial species. Of these many are probably novel and have not yet been cultured. Drinking water biofilms are well known to carry diverse microbial communities. Many of the microbes identified in this work as part of the studied biofilm are indeed very closely related to typical drinking water or fresh water microbes, and their pres-

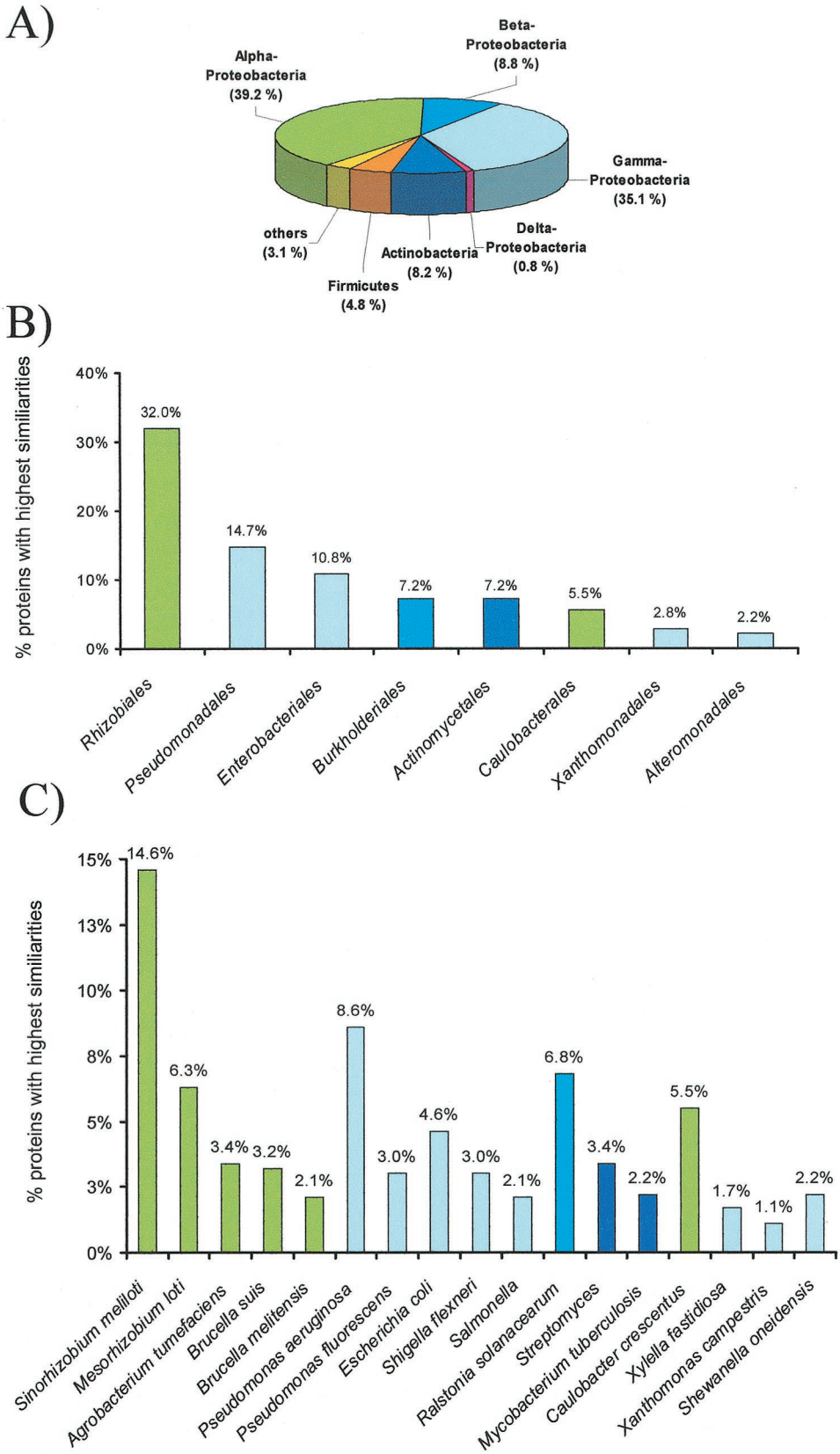


FIG. 3. Distribution of BlastX similarities among bacterial phyla (A), bacterial families (B), and bacterial genera and species (C). The results indicate the distribution of the highest similarities observed after 1,026 BlastX searches. The DNA sequences were derived from the snapshot genome sequencing project, and only those sequences which resulted in the identification of functional proteins were included. In B, only those bacterial families for which more than 20 hits (2%) could be observed were included; and in C, only the bacterial species for which more than 10 hits (1%) could be observed were included.

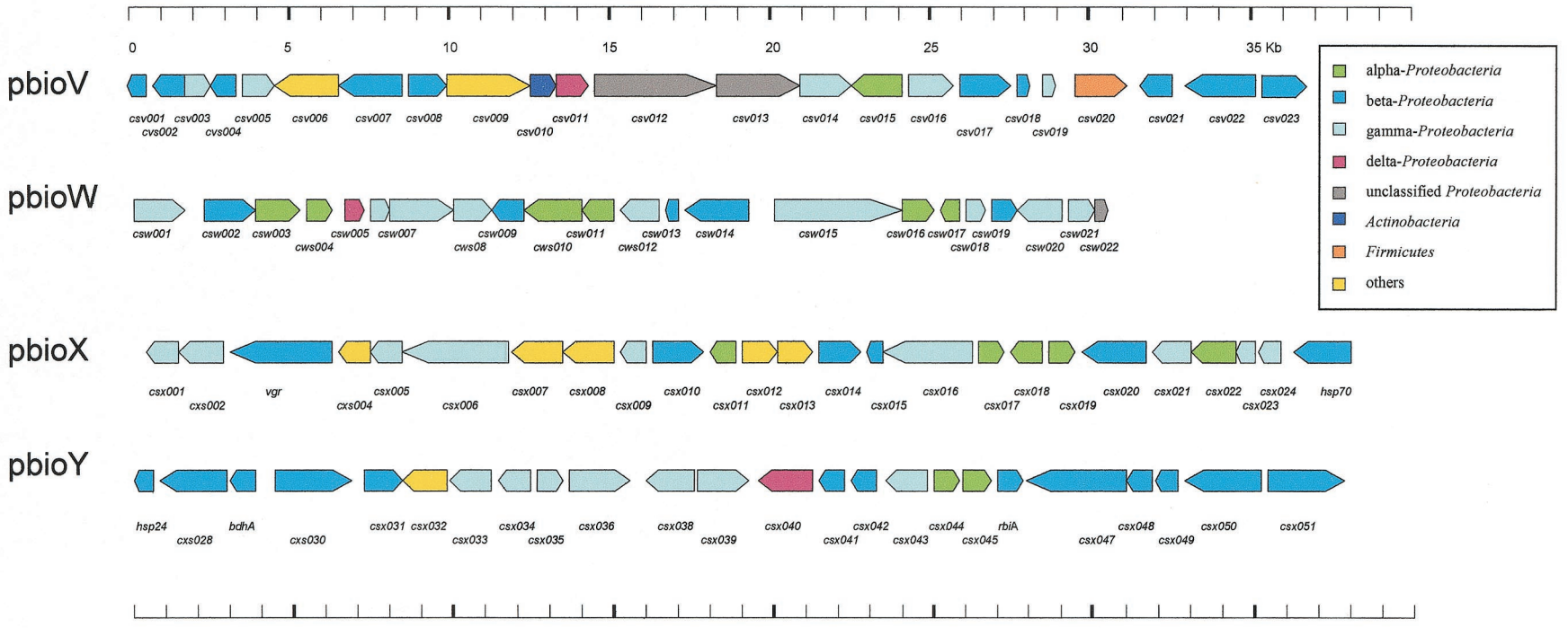


FIG. 4. Physical maps of the central parts of four cosmid clones isolated from the biofilm metagenome library. Arrows indicate the locations and directions of transcription of the identified open reading frames (ORFs) on the different cosmids. Observed similarities for the indicated ORFs are listed in Tables 4 to 6, together with the GenBank accession numbers. Color codes indicate the highest similarities of the deduced protein sequences to proteins of known bacterial species and their phylogenetic positions within the *Proteobacteria*, *Actinobacteria*, and *Firmicutes*. Only the highest similarities were considered for this analysis; color coding is identical to the color coding used in Fig. 3. The clones pbioX and pbioY form a 75-kb overlapping DNA fragment, and the DNA sequence was submitted to GenBank in two parts (contig1, *csx001* to *csx024*; contig2, *csx026* to *csx051*).

TABLE 4. Genes identified and observed similarities for ORFs identified on pBioV[a]

ORF	Size of putative protein (AA)	Function, closest match	GenBank accession no. of closest match	ORF alignment region (AA)	Alignment region (AA range/total AA)	% Identity
csv001	258	D-Pantothenate synthesis, *panC Ralstonia solanacearum*	NP520508.1	1–256	1–247/283	58
csv002	315	D-Pantothenate synthesis, *panB Neisseria meningitides*	NP283858.1	48–308	2–255/263	56
csv003	288	Sterol desaturase *Shewanella oneidensis*	AAN56787.1	15–280	11–280/287	52
csv004	265	Hypothetical protein *Burkholderia fungorum*	ZP00029113.1	2–265	28–291/291	73
csv005	324	Hypothetical protein *Pseudomonas fluorescens*	ZP00083568.1	2–315	3–317/456	29
csv006	627	Sulfate transporter *Chlorobium tepidum*	AAM71951.1	78–625	40–586/618	41
csv007	701	Methionyl-tRNA synthetase *Ralstonia solanacearum*	CAD16088.1	3–701	4–688/688	58
csv008	375	MRP ATP binding protein *Ralstonia solanacearum*	CAD16086.1	1–369	1–359/362	57
csv009	875	Histidine kinase *Nostoc* sp.	BAB73836.1	205–822	309–912/944	22
csv010	213	Response regulator *Mycobacterium tuberculosis*	AAK45932.1	10–201	18–200/208	25
csv011	357	Hypothetical protein *Geobacter metallireducens*	ZP00081330.1	113–337	181–396/397	47
csv012	1,248	Hypothetical protein *Magnetococcus sp.*	ZP0004555.1	300–1186	159–1045/1202	34
csv013	742	Hypothetical protein *Magnetococcus sp.*	ZP0043181.1	3–727	28–733/733	32
csv014	584	Hypothetical protein *Xanthomonas axonopodis*	AAM35315.1	167–580	129–537/542	30
csv015	505	Hypothetical protein *Rhodopseudomonas palustris*	ZP00012497	6–499	3–491/499	52
csv016	553	Hypothetical protein *Pseudomonas fluorescens*	ZP0084879.1	161–551	19–407/410	60
csv017	524	ABC transporter *Pseudomonas putida*	AAN67853.1	1–520	96–617/617	68
csv018	151	Hypothetical protein *Ralstonia metallidurans*	ZP00022071.1	1–150	1–150/150	66
csv019	165	Hypothetical protein *Pseudomonas syringae*	ZP00125646.1	15–164	15–165/169	37
csv020	615	*p*-Aminobenzoate-synthase component *Listeria innocua*	CAC98119.1	28–603	15–566/568	39
aroC	380	Chorismate synthase *Ralstonia solanacearum*	CAD15268.1	14–379	1–366/366	78
csv022	427	Ribonuclease *Ralstonia solanacearum*	NP519680.1	36–402	8–375/441	39

[a] AA, amino acids. DNA sequences were submitted to GenBank under accession number AY280634. Identified ORFs were designated a gene name if the observed *e* value was below 10^{-80}.

ence in biofilms has been described earlier (14, 15, 20, 21, 28, 30).

Surprisingly, many of the 16S rRNA clones analyzed in this work were highly similar to microbes closely related to rhizobial species. Microorganisms from the gram-negative genera *Rhizobium*, *Sinorhizobium*, *Bradyrhizobium*, *Mesorhizobium*, and *Azorhizobium*, collectively termed rhizobia, are well known for their capacity to establish N_2-fixing symbioses with legume plants (6). The observation here that rhizobial species or closely related microbes are possibly present within the biofilm

TABLE 5. Genes and observed similarities for ORFs identified on pBioW[a]

ORF	Size of putative protein (AA)	Function, closest match	GenBank accession no. of closest match	ORF alignment region (AA range)	Alignment region (AA range/total AA)	% Identity
csw001	491	ABC transporter *Azotobacter vinelandii*	ZP00092800.1	1–485	69–548/561	71
csw002	474	Hypothetical protein *Achromobacter xylosoxidans*	CAD24029.1	2–474	49–478/478	47
csw003	508	Hypothetical protein *Caulobacter cresentus*	AAK24385.1	5–496	3–514/524	37
csw004	215	Hypothetical protein *Rhodobacter sphaeroides*	ZP00004258.1	37–213	116–280/282	46
csw005	203	Hypothetical protein *Desulfovibrio desulfuricans*	ZP00129685.1	50–180	19–147/163	28
csw006	212	Oxireductase *Xanthomonas axonopodis*	AAM37740.1	6–212	9–222/222	67
csw007	441	Oxireductase *Xanthomonas axonopodis*	AAM37739.1	126–441	1–316/316	76
csw008	762	Oxireductase *Xanthomonas campestris*	AAM41999.1	27–759	1–730/735	73
csw009	353	Hypothetical protein *Ralstonia metallidurans*	ZP00024178.1	1–341	1–332/332	52
csw010	579	Hypothetical protein *Bradyrhizobium japonicum*	NP769440.1	5–578	7–580/580	55
csw011	270	Hypothetical protein *Bradyrhizobium japonicum*	AAO07326.1	1–267	1–288/300	74
csw012	438	Nitrate regulatory protein *Klebsiella pneumoniae*	AAA25101.2	11–425	17–389/396	31
csw013	112	Nitrite reductase (small subunit) *Ralstonia solanacearum*	NP522782.2	7–110	5–111/113	53
csw014	848	Nitrite reductase (large subunit) *Ralstonia solanacearum*	NP522783.1	1–848	3–842/852	76
csw015	1,387	DNA helicase *Xanthomonas axonopodis*	AAM37301.1	1–1374	79–1449/1480	67
csw016	278	Hypothetical protein *Rhodobacter sphaeroides*	ZP00008199.1	91–233	40–193/203	28
csw017	185	Repressor protein *Agrobacterium tumefaciens*	NP356453.1	7–183	11–187/198	53
csw018	301	LysR regulator *Pseudomonas putida*	AAN68114.1	1–299	1–299/308	64
csw019	194	Putative ligase protein *Ralstonia solanacearum*	NP519564.1	4–193	3–192/198	50
csw020	676	Oligopeptidase *Shewanella oneidensis*	AAN57578.1	43–664	19–641/645	38
csw021	320	Heme oxygenase *Pseudomonas syringae*	AAM00281.1	130–312	13–196/204	32

[a] DNA sequences were submitted to GenBank under accession number AY280635. Identified ORFs were designated a gene name if the observed *e* value was below 10^{-80}.

TABLE 6. Genes and observed similarities for ORFs identified on the 75-kb DNA fragment formed by overlapping cosmids pbioX and pbioY[a]

ORF	Size of putative protein (AA)	Function, closest match	GenBank accession no. of closest match	ORF alignment region (AA range)	Alignment region (AA range/total AA)	% Identity
csx001	327	Hypothetical protein ORF48 *Photorhabdus luminescens*	AAL18484.1	31–279	7–263/277	28
csx002	547	Possible lipase ORF47 *Photorhabdus luminescens*	AAL18483.1	1–540	2–535/537	40
csx003	1,050	Possible VgrG related protein *Ralstonia solanacearum*	CAD17919.1	35–1047	16–999/1006	39
csx004	398	Possible invertase *Arabidopsis thaliana*	AAL08305.1	23–394	134–537/617	20
csx005	350	Hypothetical protein *Azotobacter vinelandii*	ZP00092060.1	146–332	32–249/687	30
csx006	1132	Possible alpha-amylase *Pseudomonas syringae*	AAO56261.1	2–673	5–653/1108	61
csx007	390	Alpha-amylase family protein *Chlorobium tepidum*	AAM73306.1	21–341	18–337/670	47
csx008	568	1,4-Alpha-glucan branching enzyme *Aquifex aeolicus*	AAC06895.1	12–520	52–612/630	61
csx009	274	Hypothetical protein *Xanthomonas campestris*	NP636161.1	1–267	1–271/274	39
csx010	472	Proton glutamate symport protein *Ralstonia metallidurans*	ZP000247776.1	24–454	1–422/430	75
csx011	241	Transcriptional regulator *Magnetospirillum magnetotacticum*	ZP00055867.1	49–219	12–181/202	30
csx012	334	Multidrug resistance protein A *Nostoc punctiforme*	ZP001111687.1	2–319	56–382/393	37
csx013	492	Multidrug resistance protein B *Nostoc punctiforme*	ZP001111685.1	18–491	12–485/526	54
csx014	509	Multidrug resistance protein C *Burkholderia funghorum*	ZP00032009.1	12–508	5–510/516	35
csx015	212	Hypothetical protein *Burkholderia fungorum*	ZP00034179.1	1–206	1–204/210	41
csx016	867	FhuE, transport protein *Xanthomonas campestris*	AAM39477.1	125–867	1–746/746	43
csx017	268	Hypothetical protein *Rhodobacter sphaeroides*	ZP00005910.1	21–226	3–208/373	42
csx018	404	Hypothetical protein *Mesorhizobium loti*	BAB50278.1	26–402	27–403/404	58
csx019	341	Transcriptional regulator *Mesorhizobium loti*	BAB50279.1	1–326	1–319/328	55
csx020	647	Hypothetical protein *Ralstonia metallidurans*	ZP00025947.1	61–485	171–607/935	36
csx021	426	OruR regulator *Pseudomonas aeruginosa*	AAB94774.1	141–423	57–338/339	30
csx022	603	Hypothetical protein *Caulobacter crescentus*	AAK22724.1	35–583	537–1219/1245	25
csx023	184	Hypothetical protein *Pseudomonas aeruginosa*	AAG04515.1	8–184	3–187/195	31
csx024	311	Lactonizing lipase *Pseudomonas mendocina*	AAM14701	64–311	33–301/311	41
csx026	647	HSP70 *Ralstonia solanacearum*	CAD16342.1	1–644	37–684/688	83
csx027	180	HSP24 *Burkholderia fungorum*	ZP00030660.1	25–180	43–194/194	65
csx028	649	Lipoprotein *Ralstonia solanacearum*	NP519830.1	24–641	43–648/657	37
bdhA	284	Beta-hydroxybutyrate dehydrogenase *Ralstonia eutropha*	AAD33952.1	25–284	1–259/259	78
csx030	865	Pyrophosphokinase *Ralstonia solanacearum*	CAD15278.1	131–865	6–746/746	50
csx031	450	Transposase *Burkholderia fungorum*	CAD17735.1	12–411	11–438/440	66
csx032	483	Hypothetical protein/chitinase *Deinococcus radiodurans*	AAF12325.1	159–332	451–641/818	31
csx033	440	Sensor histidine kinase *Pseudomonas putida*	AAN68322.1	5–412	2–429/454	30
csx034	228	Two-component response regulator *Pseudomonas aeruginosa*	AAG08162.1	1–222	1–216/221	54
csx035	223	Hypothetical protein *Shewanella oneidensis*	AAN56949.1	9–204	22–223/264	34
csx036	615	Hypothetical protein *Pseudomonas aeruginosa*	AAG05360.1	77–613	5–538/538	43
csx038	433	Phospoglycerate dehydrogenase *Pseudomonas putida*	AAN70720.1	25–433	1–409/409	69
csx039	467	Oxidoreductase *Pseudomonas aeruginosa*	NP249008.1	13–465	10–462/464	69
csx040	499	Hypothetical protein *Geobacter metallireducens*	ZP00081903.1	6–404	6–385/638	29
csx041	213	Antioxidant peroxidase *Ralstonia solanacearum*	CAD14284.1	3–213	2–212/212	84
csx042	217	Hypothetical protein/esterase *Burkholderia fungorum*	ZP00031463.1	22–196	83–250/271	44
csx043	431	Homocysteine synthase *Xanthomonas campestris*	AAM42339.1	4–426	3–425/428	81
csx044	208	Hypothetical protein *Novosphingobium aromaticivorans*	ZP00093590.1	3–204	89–293/297	51
csx045	237	Hypothetical protein *Agrobacterium tumefaciens*	AAK87332.1	1–223	30–238/321	36
rbiA	223	Ribose 5-phosphate isomerase *Ralstonia metallidurans*	ZP00026886.1	9–223	1–211/211	62
csx047	681	Oligopeptidase A *Burkholderia fungorum*	ZP00028805.1	5–659	40–706/732	60
csx048	281	Methylenetetrahydrofolate dehydrogenase *Ralstonia solanacearum*	CAD15298.11	1–280	1–280/289	76
csx049	215	Two-component response regulator *Ralstonia metallidurans*	ZP00021742.1	5–201	53–250/254	69
csx050	860	Two-component sensor protein *Burkholderia fungorum*	ZP00028811.1	6–850	13–834/841	42
csx051	848	Pyruvate dehydrogenase E1 complex *Ralstonia eutropha*	AAA21598.1	9–823	12–821/895	63

[a] DNA sequences were submitted to GenBank under accession numbers AY2942268 (*csx001* to *csx0024*) and AY2942269 (*csx026* to *csx051*). Identified ORFs were designated a gene name if the observed *e* value was below 10^{-80}.

community is a novel finding and might suggest an ecological role for these microbes in these nutrient-deprived environments.

In the second approach applied in this work, we analyzed and evaluated the genome information of 2,496 high-quality snapshot sequences (Table 2), which encode approximately 2.0 Mb of raw DNA sequence information. We speculate that the overall biofilm metagenome of the studied drinking water biofilm has a size of at least 324 to 648 Mb. This is based on the finding that the biofilm communities of the analyzed samples consisted of more than 81 different microbial species (Fig. 2), each with a genome size of 4 to 8 Mb. Thus, the amount of genomic sequences generated corresponds to approximately 0.3 to 0.6% of the genomic information stored in the samples analyzed.

Although the available sequences do not allow a complete

analysis of the physiological and metabolic functions within this bacterial community, the sequences give a first insight into the biofilm genome structure and its metabolic potential. The genomic information suggests that the biofilm community is able to metabolize and catabolize a wide range of complex nutrients. Possible carbon sources available to the biofilm bacteria might be derived from the additives within the rubber coating, namely fatty acids, solubilizers, paraffin oils, and other compounds. However, additional experiments are necessary to correlate the occurrence and frequency of the catabolic genes identified through the snapshot sequencing with the in vivo catabolism of such compounds.

Our third strategy focused on the DNA analysis of large cosmid clones. The information on the DNA sequence has led to the identification of 94 ORFs (Fig. 4). The data obtained by whole cosmid sequencing supported the concept that our model microbial community is constructed of novel uncultured microbes closely related to *Proteobacteria*, and these findings support the data obtained through the phylogenetic analysis (Fig. 2A to D) and the snapshot sequencing analysis (Fig. 3). Although the observed similarities were surprisingly high for several of the identified genes, we have no evidence indicating from which species the sequenced cosmids were derived.

It is further noteworthy that the whole-cosmid sequencing as well as the snapshot genome sequencing did not result in the identification of genes encoding potential virulence factors. Therefore, we conclude that the microbial community within the studied microbial niche has only negligible pathogenic potential. This speculation is further supported by the phylogenetic data (Fig. 2). Although the phylogenetic analysis indicated the presence of several potentially pathogenic microbes, the majority of clones were similar to nonpathogenic microbial species.

Lastly, the sequencing data have been used to set up a publicly accessible database. Together with this information, a Blast server has been set up to allow in silico gene mining in the accumulated DNA sequences. Thus, one of the strengths of this report is that all the data generated are available in a searchable database, giving insight into the fine structure of the metagenome studied and other features of this unique biofilm community.

ACKNOWLEDGMENTS

This work was supported by the BMBF within the framework Genomforschung an Bakterien für die Analyze der Biodiversität und die Nutzung zur Entwicklung neuer Produktionsverfahren and the EU project GEMINI.

REFERENCES

1. **Amann, R. I., W. Ludwig, and K. H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143–169.
2. **Beja, O., L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong.** 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science **289:**1902–1906.
3. **Beja, O., E. V. Koonin, L. Aravind, L. T. Taylor, H. Seitz, J. L. Stein, D. C. Bensen, R. A. Feldman, R. V. Swanson, and E. F. DeLong.** 2002. Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl. Environ. Microbiol. **68:**335–345.
4. **Beja, O., E. N. Spudich, J. L. Spudich, M. Leclerc, and E. F. DeLong.** 2001. Proteorhodopsin phototrophy in the ocean. Nature **411:**786–789.
5. **Beja, O., M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong.** 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. **2:**516–529.
6. **Broughton, W. J., and X. Perret.** 1999. Genealogy of legume-*Rhizobium* symbioses. Curr. Opin. Plant Biol. **2:**305–311.
7. **Courtois, S., C. M. Cappellano, M. Ball, F. X. Francou, P. Normand, G. Helynck, A. Martinez, S. J. Kolvek, J. Hopke, M. S. Osburne, P. R. August, R. Nalin, M. Guerineau, P. Jeannin, P. Simonet, and J. L. Pernodet.** 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl. Environ. Microbiol. **69:**49–55.
8. **Entcheva, P., W. Liebl, A. Johann, T. Hartsch, and W. R. Streit.** 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl. Environ. Microbiol. **67:**89–99.
9. **Gupta, R., Q. K. Beg, and P. Lorenz.** 2002. Bacterial alkaline proteases: molecular approaches and industrial applications. Appl. Microbiol. Biotechnol. **59:**15–32.
10. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. **5:**R245–R249.
11. **Healy, F. G., R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam.** 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Appl. Microbiol. Biotechnol. **43:**667–674.
12. **Henne, A., R. Daniel, R. A. Schmitz, and G. Gottschalk.** 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. **65:**3901–3907.
13. **Henne, A., R. A. Schmitz, M. Bomeke, G. Gottschalk, and R. Daniel.** 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia col*i. Appl. Environ. Microbiol. **66:** 3113–3116.
14. **Kalmbach, S., W. Manz, and U. Szewzyk.** 1997. Isolation of new bacterial species from drinking water biofilms and proof of their in situ dominance with highly specific 16S rRNA probes. Appl Environ Microbiol. **63:**4164–4170.
15. **Kalmbach, S., W. Manz, J. Wecke, and U. Szewzyk.** 1999. *Aquabacterium* gen. nov., with description of *Aquabacterium citratiphilum sp. nov., Aquabacterium parvum* sp. nov. and *Aquabacterium commune* sp. nov., three in situ dominant bacterial species from the Berlin drinking water system. Int. J. Syst. Bacteriol. **49:**769–777.
16. **Knietsch, A., T. Waschkowitz, S. Bowien, A. Henne, and R. Daniel.** 2003. Construction and screening of metagenomic libraries derived from enrichment cultures: Generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli.* Appl. Environ. Microbiol. **69:** 1408–1416.
17. **MacNeil, I. A., C. L. Tiong, C. Minor, P. R. August, T. H. Grossman, K. A. Loiacono, B. A. Lynch, T. Phillips, S. Narula, R. Sundaramoorthi, A. Tyler, T. Aldredge, H. Long, M. Gilman, D. Holt, and M. S. Osburne.** 2001. Expression and isolation of antimicrobial small molecules from soil DNA libraries. J. Mol. Microbiol. Biotechnol. **3:**301–308.
18. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. **29:**173–174.
19. **Ochsenreiter, T., F. Pfeifer, and C. Schleper.** 2002. Diversity of Archaea in hypersaline environments characterized by molecular-phylogenetic and cultivation studies. Extremophiles **6:**267–274.
20. **Poindexter, J. S., K. P. Pujara, and J. T. Staley.** 2000. In situ reproductive rate of freshwater *Caulobacte*r spp. Appl. Environ. Microbiol. **66:**4105–4111.
21. **Ribas, F., J. Perramon, A. Terradillos, J. Frias, and F. Lucena.** 2000. The Pseudomonas group as an indicator of potential regrowth in water distribution systems. J. Appl. Microbiol. **88:**704–710.
22. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman.** 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66:**2541–2547.
23. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
24. **Schleper, C., E. F. DeLong, C. M. Preston, R. A. Feldman, K. Y. Wu, and R. V. Swanson.** 1998. Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum.* J. Bacteriol. **180:**5003–5009.
25. **Schleper, C., R. V. Swanson, E. J. Mathur, and E. F. DeLong.** 1997. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum.* J. Bacteriol. **179:**7803–7811.
26. **Schmalenberger, A., and C. C. Tebbe.** 2003. Bacterial diversity in maize rhizospheres: conclusions on the use of genetic profiles based on PCR-

amplified partial small subunit rRNA genes in ecological studies. Mol. Ecol. **12:**251–262.
27. **Schmidt, T. M., E. F. DeLong, and N. R. Pace.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. **173:**4371–4378.
28. **Schwartz, T., S. Hoffmann, and U. Obst.** 1998. Formation and bacterial composition of young, natural biofilms obtained from public bank-filtered drinking water systems. Water Res. **32:**2787–2797.
29. **Schwieger, F., and C. C. Tebbe.** 1998. A new approach to utilize PCR-single-strand-conformation polymorphism for 16S rRNA gene-based microbial community analysis. Appl. Environ. Microbiol. **64:**4870–4876.
30. **Szewzyk, U., R. Szewzyk, W. Manz, and K. H. Schleifer.** 2000. Microbiological safety of drinking water. Annu. Rev. Microbiol. **54:**81–127.
31. **Torsvik, V., and L. Ovreas.** 2002. Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. **5:**240–245.
32. **Voget, S., C. Leggewie, A. Uesbeck, C. Raasch, K. E. Jaeger, and W. R. Streit.** Prospecting for novel biocatalysts in a soil metagenome. Appl. Environ. Microbiol. **69:**6236–6242.