Journal of Biotechnology 136 (2008) 91–101

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor

Lutz Krause[a,b,*], Naryttza N. Diaz[a], Robert A. Edwards[e,f,g], Karl-Heinz Gartemann[a,c], Holger Krömeke[a], Heiko Neuweger[a], Alfred Pühler[d], Kai J. Runte[a,i], Andreas Schlüter[d], Jens Stoye[a,h], Rafael Szczepanowski[a], Andreas Tauch[a], Alexander Goesmann[a,i]

[a] *Centrum für Biotechnologie (CeBiTec), Universität Bielefeld, D-33594 Bielefeld, Germany*
[b] *Nestlé Research Center, BioAnalytical Science Department, CH-1000 Lausanne 26, Switzerland*
[c] *Lehrstuhl für Gentechnologie/Mikrobiologie, Universität Bielefeld, D-33594 Bielefeld, Germany*
[d] *Lehrstuhl für Genetik, Universität Bielefeld, D-33594 Bielefeld, Germany*
[e] *Center for Microbial Sciences, San Diego, CA, United States*
[f] *Department of Computer Science, San Diego State University, San Diego, CA, United States*
[g] *Mathematics & Computer Science Division, Argonne National Laboratory, Argonne, IL, United States*
[h] *AG Genominformatik, Technische Fakultät, Universität Bielefeld, D-33594 Bielefeld, Germany*
[i] *Computational Genomics, Centrum für Biotechnologie (CeBiTec), Universität Bielefeld, D-33594 Bielefeld, Germany*

ARTICLE INFO

Article history:
Received 15 February 2008
Received in revised form 3 June 2008
Accepted 9 June 2008

Keywords:
Biogas reactor
Methane production
Maize silage
Environmental sample
Pyrosequencing
Metagenome
Anaerobic food chain
Biofuel

ABSTRACT

A total community DNA sample from an agricultural biogas reactor continuously fed with maize silage, green rye, and small proportions of chicken manure has recently been sequenced using massively parallel pyrosequencing. In this study, the sample was computationally characterized without a prior assembly step, providing quantitative insights into the taxonomic composition and gene content of the underlying microbial community. *Clostridiales* from the phylum *Firmicutes* is the most prevalent phylogenetic order, *Methanomicrobiales* are dominant among methanogenic archaea. An analysis of Operational Taxonomic Units (OTUs) revealed that the entire microbial community is only partially covered by the sequenced sample, despite that estimates suggest only a moderate overall diversity of the community. Furthermore, the results strongly indicate that archaea related to the genus *Methanoculleus*, using CO_2 as electron acceptor and H_2 as electron donor, are the main producers of methane in the analyzed biogas reactor sample. A phylogenetic analysis of glycosyl hydrolase protein families suggests that *Clostridia* play an important role in the digestion of polysaccharides and oligosaccharides. Finally, the results unveiled that most of the organisms constituting the sample are still unexplored.

1. Introduction

Formation of biogas (methane) through fermentation of organic material in anaerobic reactors has the potential of tackling important, contemporary challenges: the degradation of biomass and production of energy from renewable primary products and wastes. Biogas is a clean renewable energy that promises to be a good substitution for traditional fossil fuels. Methane bioproduction is a complex, multi-step process, involving many different microbial species. Fermenting bacteria hydrolyze complex organic compounds, including polysaccharides, cellulose and xylan into oligomers and monomers (Schink, 1997). The produced intermediates are further transformed into acetate, carbon dioxide, and hydrogen by secondary fermenters. The final methanogenesis is conducted by methanogenic archea, which are highly specialized and can only use acetate, H_2, CO_2, formate or some C_1 compounds as energy substrates (Thauer, 1998).

To improve methane yield from biomass fermentation, a better understanding of the microbial community composition and metabolic processes carried out by microbes residing in biogas reactors is required. A total community DNA sample from an agricultural biogas reactor continuously fed with maize silage, green rye, and small proportions of chicken manure has been sequenced (Schlüter et al., 2008, this issue) by means of the massively parallel pyrosequencing (Margulies et al., 2005). Initial analysis by Schlüter et al. focused on the characterization of contigs assembled from the sample. Herein, to obtain a quantitative picture of the taxonomic composition, the entire sample was characterized without a prior assembly step. Additionally, several protein families were

* Corresponding author at: Nestlé Research Center, BioAnalytical Science Department, CH-1000 Lausanne 26, Switzerland. Tel.: +41 21 785 8144.
E-mail address: Lutz.Krause@rdls.nestle.com (L. Krause).

0168-1656/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2008.06.003

studied in more detail in order to obtain a better understanding of the bioconversion process occurring in the studied bioreactor. In particular, to scrutinize which microorganisms digest polysaccharides and oligosaccharides, the taxonomic origin of glycosyl hydrolase protein families were predicted using a phylogenetic analysis.

To quantitatively characterize the taxonomic composition of the biogas reactor sample, the source organisms or taxonomic origins of reads were inferred using two independent approaches: Fragments of 16S rDNA genes were identified and subsequently classified into a higher order taxonomy using the RDP rRNA Classifier (Wang et al., 2007). This approach has the advantage of yielding a high accuracy (between 99% for phylum and 83% for genus for 200 bp fragments), but only a limited number of reads can be taxonomically characterized. In order to obtain a more detailed picture of the composition, the second approach inferred the taxonomic origins of all Pfam protein family members identified in a sample using our recently published CARMA software (Krause et al., 2008).

Following the taxonomic analysis, the genetic potential of the microbes constituting the sample was characterized. For this task, protein encoding sequences (coding sequences, CDSs) were identified based on a search for protein family members using Pfam profile hidden Markov models (pHMM) (Finn et al., 2008). Pfam is a comprehensive collection of protein families, mainly representing protein domains. The major strengths of the Pfam-based analysis is the high accuracy of the pHMMs for the detection of short functional sequences when compared to pair-wise sequence comparison methods, such as Blast (Altschul et al., 1990). The high accuracy makes Pfam profile HMMs particularly adequate for the analysis of short DNA fragments obtained by pyrosequencing. On the other hand, one limitation of this approach is that usually only between 10% and 15% of reads from a sample match a Pfam family (Krause et al., 2008). Therefore, gene fragments were additionally identified based on a Blast comparison of reads with public protein database, including the database of Clusters of Orthologous Groups of proteins (COG) (Tatusov et al., 1997).

In this study, fragments of genes identified in community sequence reads are defined as environmental gene tags (EGTs). After assigning a putative gene function to each EGT, the resulting profiles could be used for a quantitative gene content analysis in order to reveal habitat-specific genetic fingerprints.

Using the described approach, detailed insights into the taxonomic composition and gene content of a methane-producing microbial community from an agricultural biogas plant were obtained.

2. Materials and methods

2.1. Origin and sequencing of the methane producing bacterial community of a biogas reactor

The biogas-producing microbial sample studied herein was taken from the first fermenter of the agricultural biogas plant in Bielefeld-Jöllenbeck (Germany) in August, 2007. The reactor had approximately 41 °C with a pH of 7.7 and was continuously fed with maize silage (63%), green rye (35%), and small proportions of chicken manure (2%). The sample was sequenced in a whole-genome-shotgun approach using the Genome Sequencer FLX system (Margulies et al., 2005). The sequencing run yielded 616,072 reads with an average size of about 230 nucleotides. In total 141,664,289 sequenced bases were obtained. More details concerning the isolation and sequencing of the biogas reactor sample are given in Schlüter et al., 2008, this issue.

2.2. Phylogenetic analysis using 16S rDNA anchors

The microbial composition of the biogas reactor sample was characterized by using fragments of 16S rDNA genes (in the following called 16S rDNAs) as phylogenetic anchors. 16S rDNAs were detected in a Blast search of all sample reads versus the ARB rRNA database (Ludwig et al., 2004). All sub-regions of reads having a Blast hit with E-value $< 10^{-6}$ were phylogenetically classified using the RDP rRNA Classifier, a naïve Bayesian rRNA classifier described by Wang et al. (2007).

The RDP classifier predicted the taxonomic origin of 16S rDNAs up to the rank of genus. Potential source species were inferred as follows: First, 16S rDNAs were compared to all those 16S rRNAs from the ARB database that have a known taxonomic origin (using BlastN with an E-value cut-off of 10^{-12}). Each 16S rDNA was subsequently aligned to the top matching database sequences (maximal 350) using the Muscle multiple alignment tool (with-parameters). 16S rDNAs without a sufficient similarity to any of the ARB sequences (sequence-identity $\geq$97%) were excluded from further analysis. Species were assigned to each remaining rDNA according to its matching database sequence with the highest sequence-identity.

2.3. Phylogenetic classification of Pfam protein family members

Using our recently published phylogenetic algorithm CARMA (Krause et al., 2008), the taxonomic origins of Pfam protein family members found in the biogas reactor were predicted. In the first step a similarity search of each sample read was conducted against Pfam's underlying sequence database using BlastX with the '-w 15' frameshift option. This computed the six-frame translations, predicted frameshifts, and identified candidate members of Pfam families. For this filter step, a relaxed E-value cut-off was applied: all reads without a Blast hit of E-value ≤ 10 were excluded from further analysis. Subsequently, each of the remaining translated reads was aligned to a multiple alignment containing all known members of its hitting protein family using the Pfam profile hidden Markov models. In the following, regions of reads matching a Pfam family are called environmental gene tags (EGTs).

Finally, a phylogenetic tree was reconstructed for each matching Pfam family. Found EGTs were classified into a higher-order taxonomy based on their phylogenetic relationships to family members with known taxonomic affiliations. Even for fragments as short as 100 bp, the method exhibits a high accuracy with a specificity ranging from 97% (superkingdom) to 68% (genus) (Krause et al., 2008).

2.4. Estimating the species richness

To estimate the (species) richness of the biogas reactor microbial community, identified 16S rDNAs were grouped into Operational Taxonomic Units (OTUs) in a similar approach as described in Sogin et al., 2006. First, 16S rDNAs were compared with the entire ARB database using BlastN with an E-value cut-off of 10^{-12}. Each 16S rDNA was subsequently aligned to the top matching database sequences (maximal 350) using the Muscle multiple alignment tool (with −diags and −maxiters 2 parameters). 16S rDNAs without a sufficient similarity (sequence-identity $\geq$97%, aligned region > 100 bp) to any of the best matching sequences from ARB were excluded from further analysis. All those remaining 16S rDNAs sharing a sequence-identity of at least 97% to the same reference sequence from the ARB database were grouped into one OTU using a heuristic approach. The first 16S rDNA was grouped into an initial OTU, the top matching sequence from ARB was selected as the reference of this OTU. Each of the remaining 16S rDNAs was added