Digging into the extracellular matrix of a complex microbial community using a combined metagenomic and metaproteomic approach

M. Albertsen, A. Stensballe, K. L. Nielsen and P. H. Nielsen

ABSTRACT

Knowledge about identity and function of extracellular polymeric substances (EPS) in complex microbial communities is sparse, although these components have a large influence on the function of the microbial communities. We investigated the presence of selected genes potentially involved in EPS production in a 145 Mbp metagenome prepared by Illumina sequencing from a full-scale wastewater treatment plant carrying out enhanced biological phosphorus removal (EBPR). A range of genes involved in alginate production was identified and assigned mainly to bacteria from the phylum Bacteroidetes. Furthermore, several proteins in the EPS matrix were extracted, purified and identified by mass spectrometry. By using the metagenome as a reference for the metaproteomic analysis, more proteins were identified compared to using only publicly available databases. This illustrates the low degree of similarity between the bacteria in the EBPR community and the sequenced bacteria in the public databases. Hence, the combination of metagenomics and metaproteomics presented here is needed to investigate the identity of the proteins in the EPS matrix.

Key words | activated sludge, EBPR, exopolymers, extracellular proteins, metagenomics, metaproteomics

INTRODUCTION

Extracellular polymeric substances (EPS) are a large and important part of microbial aggregates such as biofilms and flocs. They determine the microenvironments of the microorganisms and a number of chemical and physical properties of the aggregates. A large number of different polymeric components form the EPS matrix, primarily various polysaccharides, proteins, lipids and nucleic acids (Sutherland 2001; Flemming & Wingender 2010). Different species generally produce different EPS components, which to some extent determine the properties of the aggregate – meaning that the species composition of a certain mixed community to a large extent will determine the properties in relation to adhesion, aggregate strength and water binding properties (Larsen et al. 2007, 2008a; Dominiak et al. 2011a, b). This is important in many engineered systems since the structure of the microbial communities will affect biofilm formation in pipelines or on membranes, floc formation and clarification in wastewater treatment plants, and drainage/dewatering of surplus biomass.

Little is known about the composition of EPS in mixed microbial communities, primarily due to the predominance of uncultured and unidentified bacteria. However, the development of new methods in microbial ecology such as community genomics (metagenomics) (Tringe et al. 2005; Martín et al. 2006; Qin et al. 2010) and postgenomics techniques, such as proteomics (metaproteomics) (Park et al. 2008; Wilmes et al. 2008) can potentially be used to investigate the EPS components in a certain mixed microbial ecosystem.

In this study, the metagenome of a microbial ecosystem from a wastewater treatment plant carrying out enhanced biological phosphorus removal (EBPR) was used for searching for potential extracellular proteins or proteins involved in production of extracellular polysaccharides. It was chosen to focus on enzymes involved in the production of
the polysaccharide alginate and proteins forming extracellular amyloidic fibrils, both potentially present in wastewater systems and involved in floc formation (Larsen et al. 2008b; Lin et al. 2010). Furthermore, using metaproteomics, an identification of extracellular proteins extracted from the EPS matrix from the EBPR plant was also carried out.

Methodological approach

The metagenome from a mixed microbial community, e.g. an EBPR plant, can be obtained after total community DNA extraction and sequencing. The sequenced pieces of DNA (reads) are typically assembled to contigs (contiguous overlapping pieces of DNA) (Figure 1). These can be annotated using automated services as e.g. MG-RAST (Meyer et al. 2008) or local BLAST searches against e.g. the National Center for Biotechnology Information (NCBI) nr database. However, as many proteins have unknown or only very general functional annotations, many interesting proteins are often not observed by the broad search strategies. Another possibility is to generate manually curated databases that can be used for investigation of specific functions (e.g. Hess et al. 2011). After searching the metagenome against the specific database the hits can be searched against a non-specific database (e.g. NCBI nr) and the significance of the hit in the specific database can be evaluated. This enables the decision if it is indeed a protein of interest or if a (much) better hit to another protein is evident. However, it is often needed to construct phylogenetic trees in order to infer a putative phylogenetic relationship.

As the number and diversity of identified EPS-related proteins presently is low, another possibility is to increase the knowledge base through the combined use of metagenomics and proteomics of the protein fraction of the EPS layer. Proteomics has the advantage that only expressed proteins are analyzed, i.e. not the total genetic potential of the population, and is thus the functional extension of the metagenome. However, in order to utilize proteomics the generated mass spectra have to be searched against a database using e.g. the MASCOT algorithm.
This generally requires databases of proteins of very closely related organisms (>90% protein similarity; VerBerkmoes et al. 2009). However, this is rarely possible in mixed microbial communities due to the lack of suitable reference genomes from the system, e.g. from a wastewater treatment plant (Albertsen et al. 2012). The combined use of metagenomics and proteomics analysis from the same environment will enhance the possibility of identifying EPS-related proteins as we will demonstrate in this study.

MATERIALS AND METHODS

Sampling

An activated sludge sample was taken from the aeration tank at Aalborg East wastewater treatment plant, Denmark (57.044565 N, 10.047598 E) 04/08/2009. The plant has a biodenipho configuration, which includes an anaerobic tank and a tank with alternating denitrifying anoxic and nitrifying oxic conditions. The plant treats mainly domestic wastewater and serves 100,000 person equivalents (PE) with an average load of 45,000 PE. The plant has had stable EBPR activity for several years and also at the time of sampling.

Metagenomic DNA extraction, sequencing and assembly

DNA was extracted using the FastDNA® SPIN kit for soil (MP Biomedicals, Solon, OH, USA). Paired End Sequencing (2 × 72nt) was performed using an Illumina Genome Analyzer II (Illumina Inc., CA, USA). The metagenome reads were assembled using CLCs de novo assembly algorithm (CLC bio, Aarhus, Denmark), and uploaded to the MG-RAST server (Meyer et al. 2008) (accession number: 4463936). Details can be found in Albertsen et al. (2012).

Identification of EPS-related proteins

Selected proteins involved in EPS formation in various bacteria (true extracellular proteins and proteins (enzymes) forming polysaccharides) were found through a literature study. For each selected protein a keyword search in the NCBI Refseq database was used to identify homologous proteins. To remove redundant proteins each set of proteins was clustered at 95% amino acid identity using cd-hit (Li & Godzik 2006). The set of non-redundant proteins was used to create a BLAST database using the makeblastdb tool. Open reading frames (ORFs) were predicted in the metagenome using MetaGeneAnnotator (Noguchi et al. 2008) with the -m parameter for multiple species. The predicted ORFs were translated to proteins using the NCBI translation parameter for multiple species. The predicted metagenome proteins were searched against the database of EPS-related proteins using BLASTP with an e-value cut-off of 1 × 10⁻5. The resulting list of metagenome proteins predicted to be involved in EPS formation was then searched against the full NCBI nr database and proteins with better hits to proteins that were not involved in EPS formation were removed. MEGAN (Huson et al. 2007) was used to taxonomically bin the resulting proteins based on their lowest common ancestor (LCA) algorithm. A 10% bit score difference was used to assign proteins uniquely to specific nodes.

EPS protein extraction

Protein extraction from activated sludge was performed as described by Frølund et al. (1996) and Gessesse et al. (2003) with minor modifications. The use of baffled beaker was omitted and a glass dounce homogenizer was used in combination with vigorous shaking. Rod sonication was not applied to avoid rupture of bacteria. Activated sludge was cooled immediately to 4 °C followed by protein extraction facilitated by addition of 50 mM disodium EDTA and 0.1% TX-100. The crude protein extract from EPS was precipitated by chloroform/methanol followed by separation by 4–16% SDS PAGE and visualization by the EMBL silver-stain directly compatible with in-gel digest and mass spectrometry (MS) analysis (Shevchenko et al. 1996). Samples for protein identification by tandem mass spectrometry (MS/MS) were digested in situ by trypsin and analyzed using a nanoLC-QTOF MS. The MS/MS data were searched using Matrixscience Distiller De Novo based MS/MS ion search (Matrix Science, UK) against the UniProt (Eubacteria only; 322,372 sequences) or the in-house protein database (Metagenome Wastewater Treatment Plant East and West; 1,213,900 sequences; 143,957,419 residues). Semi-tryptic enzyme specificity was allowed with two missed cleavages. The peptide and fragment ion mass accuracy were set to 0.1 Da and the experimentally obtained mass precision was within 10 ppm. All peptide tandem mass spectra used for protein identification were fully tryptic and manually verified followed by validation of bacterial origin by PSI-BLAST.
RESULTS AND DISCUSSION

Identification of EPS-related proteins in the metagenome

The metagenome from Aalborg East EBPR wastewater treatment plant was established after extraction and sequencing. It consists of 18.2 Gb short read sequences that were assembled into 269,835 contigs of at least 300 bp in length (average 540 bp, maximum 32,844 bp), and from these 338,863 ORFs were predicted. The total metagenome size encompass 145 Mb of non-redundant sequence. The initial annotation of metagenome ORFs (through MG-RAST) identified 79 proteins potentially involved in alginate metabolism and 68 proteins involved in general exopolysaccharide production, thereby confirming that at least some potential for alginate production and modification is present in the community (Table 1).

The alginate pathway and several proteins directly involved in EPS formation were chosen for further studies. The chosen reference proteins were searched against the NCBI nr database and manually inspected to remove false positives (Table 2). In general, few of the identified putative EPS proteins were best hits and the percent identity of the proteins compared to the reference genomes were in the range of 30–65%. The proteins identified were taxonomically binned (the potential identity of the bacteria producing the protein) by the use of MEGAN’s LCA algorithm. Four alginate lyase genes could be binned to the phylum Bacteroidetes and one to Alphaproteobacteria. This is interesting because alginate-producing bacteria are often considered to be pseudomonads and other Proteobacteria, but since Bacteroidetes is abundant in this EBPR ecosystem (Kragelund et al. 2008; Nielsen et al. 2010), they may be the dominant alginate producers. Alternatively, they may also primarily produce alginate-degrading enzymes such as lyases as this group of bacteria are known as degraders of various macromolecules (Xia et al. 2008). However, in general it was difficult to bin the sequences below class level and most sequences were binned as ‘Bacteria’ which can be attributed to the low degree of similarity between the current metagenome and the reference genomes in the databases. The finding of proteins that are potentially involved in the periplasmic scaffolding and export in addition to a large number of proteins involved in alginate modification support the potential for alginate production in the EBPR plant.

A number of proteins have no enzymatic function but are excreted as structural elements into the EPS matrix. Among these proteins are functional amyloids that can form extracellular fibrils of importance for adhesion and biofilm/aggregate formation (Nielsen et al. 2011). Examples are the proteins FapC in Pseudomonas and CsgA in Escherichia coli. However, we did not observe the fapC or csgA genes in the metagenome, and it is likely due to the low homology of these genes (35–40% amino acids, aa) between different species of

Table 1 | Overview of proteins identified to be involved in alginate metabolism and exopolysaccharide biosynthesis through MG-RAST. For each enzyme the EC (Enzyme Commission) number is given, as well as average e-values, the average identity compared to the reference sequences

<table>
<thead>
<tr>
<th>Gene Description</th>
<th>#Proteins</th>
<th>Avg. e-value</th>
<th>Avg. %identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alginate metabolism</td>
<td>AlgC Phosphomannomutase (EC 5.4.2.8)</td>
<td>34</td>
<td>–31</td>
</tr>
<tr>
<td>AlgJ,F</td>
<td>Poly(b-D-mannuronate) O-acetylase (EC 2.3.1.-)</td>
<td>18</td>
<td>–29</td>
</tr>
<tr>
<td>AlgD</td>
<td>Man-1-P GDP-transferase (EC 2.7.7.22)</td>
<td>17</td>
<td>–22</td>
</tr>
<tr>
<td>AlgA</td>
<td>Mannose-6-P isomerase (EC 5.3.1.8)</td>
<td>6</td>
<td>–27</td>
</tr>
<tr>
<td>AlgL</td>
<td>Poly(b-D-mannuronate) lyase (EC 4.2.2.3)</td>
<td>2</td>
<td>–25</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>79</td>
<td>–26</td>
</tr>
<tr>
<td>Exopolysaccharide biosynthesis</td>
<td>Glycosyl transferase, group 2 family protein</td>
<td>42</td>
<td>–29</td>
</tr>
<tr>
<td>Undecaprenyl-p. transferase (EC 2.7.8.6)</td>
<td>10</td>
<td>–33</td>
<td>72</td>
</tr>
<tr>
<td>Manganese-dep. phosphatase (EC 3.1.3.48)</td>
<td>5</td>
<td>–10</td>
<td>65</td>
</tr>
<tr>
<td>Glycosyl transferase, group 1 family protein</td>
<td>4</td>
<td>–24</td>
<td>63</td>
</tr>
<tr>
<td>CpsA</td>
<td>Capsular polysaccharide synthesis enzyme</td>
<td>2</td>
<td>–31</td>
</tr>
<tr>
<td>CpsD</td>
<td>Tyrosine-protein kinase (EC 2.7.10.2)</td>
<td>2</td>
<td>–11</td>
</tr>
<tr>
<td>EpsF</td>
<td>Exopolys. Biosyn. Glycosyltransferase (EC 2.4.1.-)</td>
<td>1</td>
<td>–9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>68</td>
<td>–21</td>
</tr>
</tbody>
</table>
from the amyloids is that this class of proteins is much more conserved throughout the bacterial taxa and can therefore more easily be identified by bioinformatic tools. The results clearly show a large potential for the expression of these extracellular proteins by a range of bacteria in the EBPR community.

### Extraction and identification of extracellular proteins

Among the proteins extracted from the EPS matrix of activated sludge from the EBPR plant, only eight protein candidates were above the significance threshold using the metagenome as ‘database’ (Table 4). However, by using the general UniProt Eubacteria database only two proteins could be identified. The increase in the number of significant protein hits emphasizes the feasibility of combining proteomics with metagenomics to obtain a sample-specific reference for peptide matching. The primary protein candidate showed similarity with secreted proteins from *Pseudomonas* species but little is known about its potential function. Other proteins were produced by bacteria related to *Dechloromonas*, Bacteroidetes, Chloroflexi and others, which are all common in the activated sludge ecosystem (*Albertsen* et al. 2012). The low number of EPS proteins identified seemed to be related to difficulties in the extraction of true EPS proteins. Despite several attempts and different optimization procedures, the peptide yield was low after trypsin digestion. However, the protein content in the crude EPS extract was large, so future optimization...
of the analyses may reveal several other extracellular proteins.

CONCLUSION

The potential for alginate production and modification was identified in the full-scale EBPR metagenome and four of 12 potential alginate lyase genes could be binned to the phylum Bacteroidetes. The combination of metagenomics and metaproteomics made it possible to identify EPS proteins extracted directly from the complex community. Further improvement of the protein extraction method has great promise to increase the number of identifiable EPS proteins, thereby extending the knowledge of EPS proteins in mixed bacterial communities.

ACKNOWLEDGEMENTS

The study was funded by the Danish Council for Strategic Research (Centre for Design of Microbial Communities in Membrane Bioreactors).

REFERENCES


Table 4 | Overview of proteins and coding sequences (CDS) identified based on true-EPS extraction. For each protein identification the Protein MOWSE Score (number of assigned peptides), as well as the taxonomy, compared to the reference sequences is shown

<table>
<thead>
<tr>
<th>Database</th>
<th>Protein identification</th>
<th>MOWSE score (unique peptides)</th>
<th>Taxonomic binning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniprot</td>
<td>Aspartyl-RNA synthetase</td>
<td>59 (2)</td>
<td>Cyanobacteria</td>
</tr>
<tr>
<td></td>
<td>Oxidoreductase</td>
<td>55 (2)</td>
<td>Gammaproteobacteria</td>
</tr>
<tr>
<td>Metagenomic</td>
<td>Hypothetical protein</td>
<td>98 (3)</td>
<td>Pseudomonas</td>
</tr>
<tr>
<td></td>
<td>Outer membrane porin</td>
<td>73 (1)</td>
<td>Dechloromonas</td>
</tr>
<tr>
<td></td>
<td>Iojap-like protein</td>
<td>62 (2)</td>
<td>Bacteroidetes</td>
</tr>
<tr>
<td></td>
<td>Secretion target repeat protein</td>
<td>59 (1)</td>
<td>Bacteria</td>
</tr>
<tr>
<td></td>
<td>ABC transporter</td>
<td>58 (1)</td>
<td>Betaproteobacteria</td>
</tr>
<tr>
<td></td>
<td>Hypothetical protein</td>
<td>56 (2)</td>
<td>Desulfovibacterium sp.</td>
</tr>
<tr>
<td></td>
<td>NTPase</td>
<td>53 (2)</td>
<td>Bacteroidetes</td>
</tr>
<tr>
<td></td>
<td>Glycosyltransferase group 1</td>
<td>52 (1)</td>
<td>Chloroflexaceae</td>
</tr>
</tbody>
</table>


First received 3 March 2012; accepted in revised form 26 November 2012