1   **Metaproteomics of activated sludge from a wastewater treatment plant- a pilot study**

2   "Dataset Brief"

3   *Sebastian Püttker[1,*], Fabian Kohrs[1,*], Dirk Benndorf[1], Robert Heyer[1], Erdmann Rapp[2] and Udo Reichl[1,2]*

4   [1] Bioprocess Engineering, Otto von Guericke University Magdeburg, Magdeburg, Germany

5   [2] Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg,

6   Germany

7   [*] Equal authorship

8   **Correspondence**: Dr. rer. nat. Dirk Benndorf, Bioprocess Engineering, Otto von Guericke University

9   Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany. E-Mail: benndorf@mpi-

10   magdeburg.mpg.de. Fax: ++49 / 391 / 67-11209.

11   **Abbreviations**: **M**, molar; **WWTP**, wastewater treatment plant; **S**, summer; **W**, winter; **3D**, three

12   dimensional; **IAA**, iodoacetamide; **FA**, formic acid; **MGF**, Mascot generic format; **MPA**,

13   MetaproteomeAnalyzer; **KEGG**, Kyoto Encyclopedia of Genes and Genomes; **KO**, KEGG Orthology; **EC**,

14   Enzyme Commission

15   **Keywords**: Metaproteomics / Wastewater treatment plant / Activated sludge / MetaProteomeAnalyzer /

16   Offgel isoelectric focusing

17   **3,080 Words**

18   **Abstract**

19   In this study the impact of protein fractionation techniques prior to LC/MS analysis was investigated on

20   activated sludge samples derived at winter and summer condition from a full-scale wastewater

21   treatment plant (WWTP). For reduction of the sample complexity, different fractionation techniques

22   including RP-LC (1D-approach), SDS-PAGE and RP-LC (2D-approach) as well as RP-LC, SDS-PAGE and liquid

23   IEF (3D-approach) were carried out before subsequent ion trap MS analysis. The derived spectra were

24   identified by MASCOT search using a combination of the public UniProtKB/Swiss-Prot protein database

25   and metagenome data from a WWTP (data are available via ProteomeXchange with identifier

26   PXD001547). The results showed a significant increase of identified spectra, enabled by applying IEF and

27   SDS-PAGE to the proteomic workflow. Based on meta-proteins, a core metaproteome and a

28   corresponding taxonomic profile of the wastewater activated sludge were described. Functional aspects

29   were analyzed using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway library by plotting

30   KEGG Orthology identifiers (KO numbers) of protein hits into pathway maps of the central carbon

31   (map01200) and nitrogen metabolism (map00910). Using the 3D-approach, most proteins involved in

32   glycolysis and citrate cycle and nearly all proteins of the nitrogen removal were identified, qualifying this

33   approach as most promising for future studies.

34 **Text Body**

35 Due to anthropogenic processes, a high amount of organic substances like sewage, fertilizers and

36 detergents from municipal and industrial sources are released into wastewater. Discharging of polluted

37 water into the environment without further treatment endangers public health and can cause

38 eutrophication or poisoning of aquatic ecosystems. Therefore, the separation of solid material and the

39 removal of organic and inorganic contaminants in wastewater treatment plants (WWTP) are essential to

40 guarantee water purity. The biological treatment is performed by a complex community of

41 microorganisms commonly referred to as activated sludge and comprises the degradation of organic

42 material, the removal of nitrogen by nitrification and denitrification as well as the elimination of

43 phosphorus by phosphate accumulating bacteria. However, the microbial population in activated sludge

44 is sensitive to external influences. In particular, contamination with toxic compounds from industry,

45 dilution of wastewater by strong rainfall or formation of layers of scum due to seasonal changes in

46 temperature can lead to unstable operational conditions.

47 Molecular biological approaches (i.e. community fingerprinting, metagenome sequencing) revealed that

48 only a small portion of the microbial community typically found in WWTPs is cultivable in the lab. In

49 order to understand its function without isolating single strains, metaproteomic analysis has been

50 applied to activated sludge samples [1, 2]. Major challenges for this approach are the presence of humic

51 compounds interfering with protein separation, the high complexity of the community, and the lack of

52 sufficient genomic sequences for protein identification [1, 3].

53 In this work a proteomic workflow previously established for digestates of biogas plants [4] was applied

54 to characterize composition of microbial communities of activated sludge from WWTP. Different gel-free

55 and gel-based fractionation techniques were performed, to achieve a high resolution of low abundant

56 proteins and to enable detailed description of the complex metaproteome. Activated sludge was

57 sampled in winter (W, 21st of January 2014) and summer (S, 2nd of June 2014) period to cover the core

58   metaproteome as well as to represent major seasonal changes. Samples were collected from three

59   different spots of the aerobic basin of the WWTP Gerwisch (Saxony-Anhalt, Germany) and handled as

60   triplicates (see Supplementary Table S1 for general process data). A detailed description of the methods

61   used is given in Supplementary Note 1. Briefly, cell disruption and protein extraction were performed

62   simultaneously using a ball mill with liquid phenol (10 g phenol plus 1 ml water). Proteins were

63   precipitated afterwards with ammonium acetate in methanol (0.1 M) and washed with ice-cold 80%

64   acetone and 70% ethanol. Finally, extracts were suspended in 7 M urea/2 M thiourea buffer containing

65    $0.01\,g\,mL^{-1}$ DTT and protein contents were determined using the amido black dye assay. For

66   fractionation of the protein samples, SDS-PAGE, gel-free liquid IEF (allowing the recovery of proteins

67   after isoelectric focusing) and RP-LC were used and combined (1D: RP-LC; 2D: SDS-PAGE + RP-LC; 3D: IEF

68   + SDS-PAGE + RP-LC) as published previously [4]. SDS-PAGE gels were cut into equal pieces for tryptic

69   digestion, resulting in ten fractions for the 2D and 120 fractions for 3D-approach, respectively. After LC

70   the samples were directly transferred to an ESI-amaZon™ iontrap MS (Bruker Daltonics, Bremen,

71   Germany) for generation of MS/MS spectra. Identification of the MS spectra was performed using

72   MASCOT 2.4 (Matrix Science, London, England). In a first round of data evaluation the public

73   UniProtKB/Swiss-Prot protein database (state 2014/10/23, [5]) has been used for protein identification,

74   but were insufficient. In previous studies the usage of a metagenome database showed at least a two-

75   fold increase in protein identifications compared to public protein databases such as UniProt/Swiss-Prot

76   [2]. Obviously, a database derived by metagenome studies from the same plant should result in the best

77   identification rate and should therefore be preferred to third-party metagenome. Nevertheless, the

78   functional core microbiome is expected to be very similar at different time points and between

79   comparably operated plants, enabling the usage of third-party metagenomes. However, metagenome

80   data from the WWTP sampled in this study were not available, for which reason data, comprising

81   1,687,643 protein sequences, from the similar operated WWTP in Aalborg (Denmark, [6]) was used as an

82   alternative source. With this additional database about twice as much identifications could be achieved.

83 For data analysis the in-house software MetaProteomeAnalyzer 1.0.4 (MPA) was used for parsing

84 MASCOT .dat-files and claiming a FDR <1% [7]. Identified peptide sequences based on the metagenome

85 data were assigned from UniProt/Swiss-Prot database, using the protein basic local alignment search

86 tool algorithm (BLAST [8]) and an e-value cutoff of 0.0001.

87 The protein profiles of activated sludge samples from W and S period showed strong background

88 smearing of the Coomassie Brilliant Blue dye for all lanes after SDS-PAGE (2D-approach) (Figure 1A).

89 Humic compounds, which share similar physico-chemical characteristics as proteins, apparently could

90 not be completely eliminated during the sample preparation procedure. However, a few weak protein

91 bands at 25 kDa, 33 kDa and 43 kDa were observed for all three cascade sample points. Winter and

92 summer period samples showed no significant differences in the protein profiles but rather the

93 intensities of S period samples were slightly higher as judged by visual assessment. The combination of a

94 liquid isoelectric focusing prior to a SDS gel electrophoresis (3D-approach) led to an increased number of

95 protein bands over the pH range 3– 10 (Figure 1B). Also, the background staining with Coomassie dye

96 was slightly reduced due to the distribution of humic compounds over a wide pH range. In addition, part

97 of humic compounds precipitated in the IPG-Strip during the focusing step. Unfortunately, some regions

98 of the gel lanes still remained smeared. Some protein bands were only observed in single pH fractions

99 whereas others occurred in up to five fractions, indicating the presence of protein isoforms with

100 different pH values as previously shown in 2D-gels by Kuhn et al. (2011) [1]. The visual comparison of gels

101 from W and S period samples revealed unique bands (Figure 1B, red boxes) indicating that seasonal

102 changes were already detectable by combined liquid isoelectric focusing and SDS-PAGE.

103 In a next step, tryptic digests were analyzed by LC-MS/MS and all separation techniques quantitatively

104 compared regarding the number of identified peptides and proteins. Furthermore, proteins were

105 assigned using the MPA software. The number of peptides was characterized as total identified spectra

106 and furthermore the number of non-redundant peptides was calculated (Figure 2,A). Hereby, the 1D-

107   approach revealed a maximum of 50 identified spectra and 27 non-redundant peptides for sample W_3

108   and 68 total spectra and 44 non-redundant peptides in sample S_3, respectively. The peptides were

109   assigned to 327 (W_3) and 1,395 proteins (S_3, Figure 2B). Performing SDS-PAGE prior to LC-MS/MS (2D-

110   approach) increased the maximum number of identified spectra to 782 in sample W_1 (442 non-

111   redundant) and 722 in sample S_1 (312 non-redundant). The best 2D-approach yielded 4,366 and 2,799

112   protein identifications in sample W_1 and S_2. Obviously the 1D-approach showed unevenly distributed

113   identifications resulting in insufficient reproducibility, whereas the 2D-approach revealed more

114   consistent results.

115   For the 3D-approach liquid IEF was carried out before gel electrophoresis and LC. In comparison to the

116   previous approaches, 8,120 peptides (including 2,037 non-redundant hits) in sample W_1 and 4,613

117   peptides (1,512 non-redundant hits) in sample S_1. Peptides of both samples were matched to 10,095

118   proteins in sample W_1 and 9,483 in sample S_1 (Figure 2,B), respectively. The high number of identified

119   proteins compared to the peptides was mainly attributed to redundant assignments of peptides to highly

120   conserved proteins from different species, e.g. chaperones, elongation factors or keratins as discussed

121   by Benndorf and Reichl (2014) [3]. In order to remove redundancy, all proteins sharing at least one

122   peptide were combined to meta-proteins (Figure 2B) for the following analyses. Using the MPA software,

123   taxonomies of meta-proteins were assigned based on the taxonomy of the most specific peptide. After

124   this processing step significantly fewer meta-proteins remained for the winter sample: a top of 13 meta-

125   protein for the 1D-approach (W_3), 206 meta-proteins for the 2D-approach (W_1) and 936 meta-

126   proteins for the 3D-approach (W_1; similar trends were obtained for sample S_1 in Figure 2B).

127   Further taxonomic and functional analyses were exclusively carried out with the datasets obtained from

128   the 2D-approach. Therefore, hierarchical taxonomic profiles were produced with the Krona visualization

129   tool [9] using the spectral count of the meta-proteins along with the corresponding taxonomic

130   annotations. A Krona plot exemplarily displaying the complete results of the identified taxonomic

131   hierarchy from superkingdom to species was created for the sample W_1 (Downloadable File D1,

132   supporting material). Surprisingly, the chart is not only dominated by bacterial (61% of all spectra) but

133   also by eukaryotic meta-proteins (37%), which were mainly represented by the phylum chordata (*Mus*

134   *musculus*, *Homo sapiens*). Major proteins were keratins and the pancreatic enzyme elastase. The latter

135   one is released into the colon and was also detected previously in activated sludge samples [1, 3]. Small

136   fractions of archaeal (2%) and viral (0.4%) meta-proteins were also identified. The presence of virus and

137   phages in activated sludge was shown before by metagenomics and electron microscopy [10]. Numerous

138   fractions with unknown taxonomy, in particular for lower taxonomic levels, were observed showing that

139   not all meta-proteins could be assigned to specific organisms. This is due to peptides sharing the

140   common ancestor on a higher taxonomic level.

141   Focusing on the microbial activity during wastewater treatment, a detailed comparison of the W_1 and

142   S_1 samples (2D-approach) was carried out by considering only results of bacterial and archaeal meta-

143   proteins in a next step. Applicable taxonomic levels for comparison were superkingdom, phylum, class

144   and order (Downloadable Files D2 and D3, supporting material). The Krona plots of both samples

145   revealed a different number of spectral counts for bacteria (462 spectra in W_1; 225 in S_1).

146   Nevertheless, similarities could be observed regarding the ratios of bacteria (95-96%) and archaea (4–

147   5%). At the phylum level both samples were dominated by Proteobacteria (69% in sample W_1; 74% in

148   sample S_1), followed by Firmicutes (similar percentages in both samples and replicates; data not

149   shown). The class level revealed major fractions of Betaproteobacteria (W_1: 53%; S_1: 63%) and

150   Gammaproteobacteria (W_1: 15%; S_1: 13%) and minor fractions of Alpha- and Deltaproteobacteria,

151   showing no significant differences in activated sludge samples from W and S period. Slight variations

152   between S and W conditions at the order level were observed regarding the members of the Gamma-

153   and Betaproteobacteria, particularly Pseudomonadales, Enterobacteriales, Burkholderiales and

154   Nitrosomonadales. However, this variation could not be validated as a general trend neither by

155   evaluating the 2D replicates nor the 3D analyses due to lack of sensitivity or missing statistical proofs.

156   Subsequently, this observation has to be validated with further data. The presence of the Archaea is

157   mainly represented by Euryachaeota (92% in sample W_1, 100% in sample S_1), including the order

158   Archaeoglobales and Methanosarcinales. Archaea of the order Thermococcales, Methanococcales and

159   Halobacteriales were only identified in sample S_1.

160   A functional analysis of the identified meta-proteins was performed using the Kyoto Encyclopedia of

161   Genes and Genomes pathway database (KEGG, [11, 12]) and the KEGG Orthology (KO) identifier. The

162   KEGG pathway maps of the carbon metabolism (map01200) were used exemplarily for a general

163   overview as well as the nitrogen metabolism (map00910) as a WWTP specific metabolism, respectively.

164   Unfortunately the data of the 2D-approach showed only few identified KOs and allowed no

165   discrimination between samples of winter and summer (data not shown), possibly due to a lack of

166   sensitivity to detect low abundant proteins. However, using the data from the 3D-approach, revealed a

167   much more detailed identification of proteins involved in the carbon metabolism (map01200, Figure 3).

168   Several enzymes from glycolysis and citrate cycle were identified in W and S samples, mainly assigned to

169   members of the Proteobacteria (Alpha-, Beta- and Gammaproteobacteria). Single enzymes representing

170   the reductive pentose phosphate cycle (Calvin cycle, Enzyme Commission number (EC): 2.7.1.19;

171   4.1.1.39) were found uniquely in sample W_1, all assigned to Bacilli of the phylum Firmicutes.

172   The process of nitrification, the conversion of ammonia in two steps to nitrite by Nitrosomonadales and

173   to nitrate by Nitrospirales, plays an important role in the wastewater treatment process. For monitoring

174   nitrification, data of the 3D-approach were mapped in the KEGG pathway map for nitrogen (map00910),

175   equivalent to the carbon metabolism (Supplementary Figure S1). In sample W_1 ten proteins were

176   identified, covering not only all the enzymes involved in nitrification (ammonia monooxygenase

177   EC 1.14.99.39, hydroxylamine oxidoreductase EC 1.7.2.6, nitrite oxidoreductase EC 1.7.99.4) but also

178   most enzymes of denitrification such as nitrate, nitrite and nitrous oxide reductases (EC 1.7.99.4; 1.7.2.1;

179   1.7.2.4). The sample from summer period S_1 showed similar results but revealed only ammonia

180   monooxygenase as enzyme for nitrification.

181   In conclusion, metaproteomic data of WWTP activated sludge samples from winter and summer period

182   were fractionated with different techniques prior to LC-MS/MS analysis. The results revealed an

183   increased protein identification enabled by application of liquid IEF and SDS gel electrophoresis (2D and

184   3D approach, Figure 2). Additionally, the availability of metagenomic data showed a significant gain of

185   about two-fold peptide identifications thereby widen the potential of metaproteomics in analysis of

186   activated sludge. On 2D-level, taxonomic and functional assignment of meta-proteins revealed a core

187   metaproteome and characterized its most abundant community members, which presence is in

188   accordance with previous studies [6]. Significant seasonal changes between W and S could not be

189   observed, probably due to weak differences in the process conditions (Supplementary Table 1) or due to

190   the lack of resolving power of the 2D-technique. However, using the 3D-approach comprehensive

191   identification of the protein set involved in nitrogen removal was enabled (Supplementary Figure S1),

192   qualifying this technique as most promising for metabolic pathway analysis of complex microbial

193   communities. For defining a core metaproteome of WWTPs and discovering potential seasonal or

194   process related changes, more comprehensive studies using the 3D-approach are required.

203   *The authors have declared no conflict of interest.*

References

[1] Kuhn, R., Benndorf, D., Rapp, E., Reichl, U. *et al.*, Metaproteome analysis of sewage sludge from membrane bioreactors. *Proteomics* 2011, *11*, 2738–2744, DOI: 10.1002/pmic.201000590.

[2] Albertsen, M., Stensballe, A., Nielsen, K. L., Nielsen, P. H., Digging into the extracellular matrix of a complex microbial community using a combined metagenomic and metaproteomic approach. *Water science and technology a journal of the International Association on Water Pollution Research* 2013, *67*, 1650–1656, DOI: 10.2166/wst.2013.030.

[3] Benndorf, D., Reichl, U., Proteomics in environmental and technical microbiology. *Eng. Life Sci.* 2014, *14*, 27–46, DOI: 10.1002/elsc.201200205.

[4] Kohrs, F., Heyer, R., Magnussen, A., Benndorf, D. *et al.*, Sample prefractionation with liquid isoelectric focusing enables in depth microbial metaproteome analysis of mesophilic and thermophilic biogas plants. *Anaerobe* 2014, *29*, 59–67, DOI: 10.1016/j.anaerobe.2013.11.009.

[5] The UniProt Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 2013, *42*, D191-D198, DOI: 10.1093/nar/gkt1140.

[6] Albertsen, M., Hansen, L. B. S., Saunders, A. M., Nielsen, P. H. *et al.*, A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J* 2012, *6*, 1094–1106, DOI: 10.1038/ismej.2011.176.

[7] Muth, T., Behne, A., Heyer, R., Kohrs, F. *et al.*, The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *J. Proteome Res.* 2015, *14*, 1557–1565, DOI: 10.1021/pr501246w.

[8] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. *et al.*, Basic local alignment search tool. *J. Mol. Biol.* 1990, *215*, 403–410, DOI: 10.1016/S0022-2836(05)80360-2.

[9] Ondov, B. D., Bergman, N. H., Phillippy, A. M., Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 2011, *12*, 385, DOI: 10.1186/1471-2105-12-385.

[10] Flannery, J., Keaveney, S., Rajko-Nenow, P., O'Flaherty, V. *et al.*, Norovirus and FRNA bacteriophage determined by RT-qPCR and infectious FRNA bacteriophage in wastewater and oysters. *Water research* 2013, *47*, 5222–5231, DOI: 10.1016/j.watres.2013.06.008.

[11] Kanehisa, M., Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 2000, *28*, 27–30.

[12] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M. *et al.*, Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* 2014, *42*, D199-205, DOI: 10.1093/nar/gkt1076.

[13] Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A. *et al.*, The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* 2013, *41*, D1063-9, DOI: 10.1093/nar/gks1262.

238   **Figure Legends**

239   **Figure 1.** *SDS-PAGE gels of the 2D and 3D-approach applied for protein extracts of activated sludge of*

240   *winter (W) and summer (S) period. STD: proteins standard; (A): SDS-PAGE gels of the 2D-approach; (B):*

241   *SDS-PAGE of the 3D-approach. Red boxes indicate differences in the protein patterns of W and S.*

242   **Figure 2.** *Peptides, proteins and meta-proteins identified by LC-MS/MS approach for all replicates of*

243   *winter (W) and summer (S) period using respective separation techniques. (A): peptides identified only*

244   *once (single peptide hits) and repeatedly (multiple peptide hits); (B): proteins and meta-proteins*

245   *identified.*

246   **Figure 3.** *KEGG pathway map showing central carbon metabolism (map01200), KEGG Ontologies (KO)*

247   *and Enzyme Commission numbers (E.C.) identified by 3D-approach of activated sludge of winter (W) and*

248   *summer (S) period. Green-colored edges indicate identified KOs from sample W_1, red colored edges*

249   *sample S_1, and blue colored edges KOs present in both samples.*

250   **Supplementary Figure S1.** *KEGG pathway map showing nitrogen metabolism (map00910), KEGG*

251   *Ontologies (KO), and Enzyme Commission numbers (E.C.) identified by 3D-approach of activated sludge of*

252   *winter (W) and summer (S) period. All enzymes identified in sample W_1 were marked with green color,*

253   *all enzymes in sample S_1 with red, respectively.*

254   **Supplementary Table S1.** *Process data acquired from afflux and drain of biological treatment step.*

Process parameter	Winter (01/21/2014)		Summer (06/02/2014)	
	Afflux[1]	Drain	Afflux	Drain
Afflux volume [m³]	100,347	-	94,469	-
Temperature [°C]	12.1	13.4	15.1	16.8
pH value [-]	7.89	7.1	7.62	7.0
Oxygen concentration [mg/L]	-	2.6	-	2.4
Chemical oxygen demand [mg/L]	685.0	36.8	269.0	33.3
Biological oxygen demand [mg/(L*5d)]	241	2	208.	2
Phosphorus concentration [mg/L]	13.10	0.10	7.52	0.13
Total nitrogen concentration [mg/L]	96.30	-	34.90	-
Ammonium concentration [mg/L]	68.80	0.02	30.00	0.03
Nitrate concentration [mg/L]	0.45	6.9	0.23	6.4
Nitrite concentration [mg/L]	0.05	0.02	0.05	0.03
Dry matter [g/L]	-	3.40	-	3.25
Sludge activity [$mg_{Oxygen}/g_{Dry\ matter}$]	-	4.40	-	2.26
Sludge content [mL_{Sludge}/L]	-	340	-	340
Sludge volume index [$mL_{Sludge}/g_{Dry\ matter}$]	-	103	-	100
Total acid content [mmol/L]	9.86	4.22	4.74	4.0

255   1: Afflux data recording is carried out two-daily, therefore process parameters shown in this column are taken from January 20th.