Journal of Biotechnology 261 (2017) 24–36

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Challenges and perspectives of metaproteomic data analysis

Robert Heyer[a,*], Kay Schallert[a], Roman Zoun[b], Beatrice Becher[a], Gunter Saake[b], Dirk Benndorf[a,c,**]

[a] Otto von Guericke University, Bioprocess Engineering, Universitätsplatz 2, 39106 Magdeburg, Germany
[b] Otto von Guericke University, Institute for Technical and Business Information Systems, Universitätsplatz 2, 39106 Magdeburg, Germany
[c] Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Sandtorstraße 1, 39106, Magdeburg, Germany

ARTICLE INFO

Keywords:
Bioinformatics
Software
Big data
Environmental proteomics
Microbial communities
Mass spectrometry

ABSTRACT

In nature microorganisms live in complex microbial communities. Comprehensive taxonomic and functional knowledge about microbial communities supports medical and technical application such as fecal diagnostics as well as operation of biogas plants or waste water treatment plants. Furthermore, microbial communities are crucial for the global carbon and nitrogen cycle in soil and in the ocean. Among the methods available for investigation of microbial communities, metaproteomics can approximate the activity of microorganisms by investigating the protein content of a sample. Although metaproteomics is a very powerful method, issues within the bioinformatic evaluation impede its success. In particular, construction of databases for protein identification, grouping of redundant proteins as well as taxonomic and functional annotation pose big challenges. Furthermore, growing amounts of data within a metaproteomics study require dedicated algorithms and software. This review summarizes recent metaproteomics software and addresses the introduced issues in detail.

1. Introduction

Microorganisms represent 50–78% of Earth's total biomass (Kallmeyer et al., 2012) and occur in all environments. Some microorganisms produce biomass by photosynthesis whereas others act as composers and degrade dead biomass. Microbial species live in complex microbial communities in which they have to compete or cooperate with each other. Understanding the functioning of the microbial communities is important, because microbial communities in the human gut effect health (Erickson et al., 2012; Heintz-Buschart et al., 2016; Kolmeder et al., 2016) and several technical applications such as waste water treatment plants (Püttker et al., 2015; Wilmes et al., 2008) and biogas plants (Abram et al., 2011; Hanreich et al., 2012) rely on the metabolic activity of microbial communities.

Methods for the investigation of microbial communities target the microbial cells, their genes, their transcripts, their proteins and their metabolites (Heyer et al., 2015). Since proteins carry out most functions in cells, including catalysis of biochemical reactions, transport and cell structure, protein amounts correlate quite well with microbial activity (Wilmes and Bond, 2006). The investigation of all proteins from one species is called proteomics. In contrast metaproteomics is the study of proteins from multiple organisms. It was introduced by Wilmes and Bond (2006, 2004) and Rodriguez-Valera (2004). The typical metaproteomics workflow comprises protein extraction and purification, tryptic digestion into peptides, protein or peptide separation and tandem mass spectrometry (MS/MS) analysis. Proteins are identified by comparing experimental mass spectra and theoretical mass spectra predicted from comprehensive protein databases. For a detailed discussion about the metaproteomics workflow please refer to Hettich et al. (2013), Becher et al. (2013), Heyer et al. (2015), Wohlbrand et al., (2013). Up to now most metaproteomics studies characterize the taxonomic and functional composition of complex microbial communities in their specific environment (Abram et al., 2011; Kan et al., 2005; Ram et al., 2005; Wilmes and Bond, 2006). A few recent studies additionally correlated the taxonomic and functional composition with certain environmental/process parameters or diseases (Erickson et al., 2012; Heyer et al., 2016; Kolmeder et al., 2016). However, three issues within bioinformatic data evaluation hampered previous

Abbreviations: CPU, central processing unit; COG, clusters of orthologous groups; DBMS, database management system (DBMS); de.NBI, German Network for Bioinformatics Infrastructure; EC, enzyme commission number; eggNOG, evolutionary genealogy of genes non-supervised orthologous; FDR, false discovery rate; GPU, graphical processing unit; GO, gene ontologies; iPath, interactive pathways explorer; LC, liquid chromatography; LCA, lowest common ancestor; KO, KEGG ontologies; MPA, MetaProteomeAnalyzer; MS, mass spectrometer; MS/MS, tandem mass spectrometer; *m/z*-ratio, mass-to-charge ratio; NoSQL, not only SQL; SQL, structured query language

* Corresponding author.
** Corresponding author at: Otto von Guericke University, Bioprocess Engineering, Universitätsplatz 2, 39106 Magdeburg, Germany.
E-mail addresses: heyer@mpi-magdeburg.mpg.de (R. Heyer), kay.schallert@ovgu.de (K. Schallert), roman.zoun@ovgu.de (R. Zoun), beatrice.becher@st.ovgu.de (B. Becher), saake@iti.cs.uni-magdeburg.de (G. Saake), benndorf@mpi-magdeburg.mpg.de (D. Benndorf).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.1201
Received 15 February 2017; Received in revised form 20 June 2017; Accepted 23 June 2017
Available online 27 June 2017

metaproteomics studies (Muth et al., 2013).

First, metaproteomes consist of up to 1000 different species (Schlüter et al., 2008). Due to high complexity metaproteomics data analysis requires a greater computational effort, necessitating bigger hard drives, more memory, more processors and more efficient algorithms. A main issue is the database search against comprehensive protein databases. Whereas handling of small protein databases below 1 GB is not critical, usage of the entire NCBI reference database requires extended computational time and may fail due to software or hardware limitations.

Second, identical peptides belonging to homologous proteins cause redundant protein identification (Herbst et al., 2016). As a result taxonomic and functional interpretation of results becomes ambiguous. A peptide may belong to the lactate dehydrogenase (EC. 1.1.1.27) (1.1.1.27) of different members of the genus *Lactobacillus*, which ferment sugars to lactate. But it may also belong to some representatives of the order *Clostridiales fermenting* lactate to acetate (Kohrs et al., 2014).

Third, protein identification is difficult if the taxonomic composition is unknown or protein entries are missing from protein databases. For example the UniProt/TrEMBL database contains only proteins from 698,745 species (http://www.ebi.ac.uk/uniprot/TrEMBLstats, status 16.12.2016), but the number of microbial species on Earth is estimated to be up to one trillion (Locey and Lennon, 2016). Thereby, already small changes in the protein sequence between related microorganisms have a big impact on protein identification. One mutation in every tenth amino acid leads to completely different tryptic peptides which hinder the identification of any peptide for the investigated protein. Thus, researchers started to sequence metagenomes alongside metaproteomics studies (Ram et al., 2005; Tyson et al., 2004). Alternatively, they use metagenomes from similar samples for protein identification.

As a consequence of these issues, standard proteomics software is often insufficient for metaproteomics studies missing the identification of unsequenced species or the comprehensive taxonomic and functional description of microbial communities. Thus, researchers favor special tools. Therefore, this review provides an overview about dedicated metaproteomics software and bioinformatic strategies.

In addition to two previous reviews on bioinformatics in metaproteomics (Muth et al., 2013, 2016) we present the impact of combining metagenomes on protein identification and address future hardware requirements and handling of big data.

After a brief introduction current metaproteomics software tools are discussed. Subsequently, this review illuminates the creation of protein databases for protein identification investigating several biogas plant samples in a use case. Then the grouping of redundant protein identifications, the evaluation of taxonomic and functional results as well as quantification in metaproteomics studies are discussed. Finally, data storage and deployment solutions for big data as well as future challenges, perspectives and demand for metaproteomics software are considered.

2. Status of proteomics software and latest trends

For the comprehensive bioinformatic processing of MS data different software tools exist. These include software for peak picking in MS-spectra, software for protein identification via database search algorithms and tools for comparison of protein expression patterns. A comprehensive summary of all these software tools can be found in the OMIC tools database (http://omictools.com/, retrieved: 09-02-2017, (Henry et al., 2014)) and in several reviews (Cappadona et al., 2012; Gonzalez-Galarza et al., 2012).

Latest trends in proteomics software are the development of proteomics tool libraries such as OpenMS (Sturm et al., 2008), Compomics (Barsnes et al., 2011) or Trans-Proteomic Pipeline (Keller and Shteynberg, 2011). These libraries comprise software tools for each step of the processing workflow, ranging from data management to data analysis. Noteworthy are also webservices, such as Expasy (Gasteiger et al., 2003), which provide a collection of small bioinformatic tools for biochemical analyses of proteins.

Repositories for MS-data such as PRIDE are used to enable long-term storage and to make published MS-data available to other researchers (Vizcaino et al., 2016). In this context general formats for exchange of MS results are necessary. Current standard in the proteomics community are the mzIdentML format (Jones et al., 2012), mzTab format (Griss et al., 2014) and mzML format (Martens et al., 2011).

Recent proteomics software combines several database search algorithms. For example, the SeachGUI tool (Vaudel et al., 2011) enables the parallel protein database search with eight different database search algorithms. Further developments are software tools for improved MS-operation and quantification. Search items for these developments are “data independent acquisition” (Doerr, 2015), “multiple and single reaction monitoring” (Colangelo et al., 2013) as well as “absolute quantification” (Cappadona et al., 2012).

Within the last years many powerful software tools were developed but their use was often restricted to a few scientific groups. Reasons were missing maintenance or availability after funding periods ended. Furthermore, many biological research groups lack bioinformatic skills to set up comprehensive software workflows or client-server architectures. In some cases even the conversion of data into the required input formats fail. In order to tackle these problems governments started to fund the collection, maintenance and support of research software tools. Examples are the Galaxy project (https://usegalaxy.org/, retrieved: 09-02-2017), (Afgan et al., 2016), ELIXIR (https://www.elixir-europe.org/, retrieved: 09-02-2017, (Crosswell and Thornton, 2012)) or de.NBI (https://www.denbi.de/, retrieved: 09-02-2017).

3. Software dedicated for metaproteomics

To address the three issues specific to metaproteomics bioinformatic data evaluation, researchers started to develop special software tools and workflows [Table 1, Fig. 1]. These tools apply different concepts, which will be discussed later. Graph 2Pep/Graph2Pro (Tang et al., 2016) and Compile (Chatterjee et al., 2016) focus on tailoring protein databases for optimal protein identification. UniPept (Mesuere et al., 2015), Prophane (Schneider et al., 2011), Megan CE (Huson et al., 2016) and Pipasic (Penzlin et al., 2014) enable taxonomic analysis, functional data evaluation and/or protein grouping. Additionally, several groups assembled comprehensive software workflows for metaproteomics, e.g. Galaxy-P (Jagtap et al., 2015), MetaPro-IQ (Zhang et al., 2016), MetaProteomeAnalyzer (Muth et al., 2015a) and others (Heintz-Buschart et al., 2016; May et al., 2016; Tanca et al., 2013). Among these workflows, the MPA is particularly user-friendly. It allows the user to control the entire bioinformatic workflow via an intuitive graphical user interface. Another noteworthy metaproteomics software tool is MetaProSIP (Sachsenberg et al., 2015). It supports the detection and quantification of isotope ratios for Protein-SIP experiments.

To ensure comparability of results between all these tools, standards for data exchange are crucial (Timmins-Schiffman et al., 2017). Consequentially, the Human Proteomics Standard Initiative is planning to extend the proteomics mzIdentML format in order to support metaproteomics data. Version 1.2.0 of the mzIdentML format (Jones et al., 2012) will support the representation of redundant protein groups (http://www.psidev.info/mzidentml, retrieved: 09-02-2017).

Another often neglected aspect is the reproducibility of results using different metaproteomics software tools. So far, only Tanca et al. (2013) tested their complete metaproteomics workflow for a defined mixed culture of nine different microorganisms. A comparison where multiple research groups evaluate an identical sample would also be desirable.

4. Construction of user databases for protein identification

Protein database selection affects the number of identified proteins as well as the identified taxonomies and identification increases. In