Challenges and perspectives of metaproteomic data analysis

Robert Heyera,⁎, Kay Schallerta, Roman Zounb, Beatrice Bechera, Gunter Saakeb, Dirk Benndorf⁎⁎

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 (Puttker 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., 2013). 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

Please cite this article as: Heyer, R., Journal of Biotechnology (2017), http://dx.doi.org/10.1016/j.jbiotec.2017.06.1201

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).
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) (1.1.1.1.27) 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 (Locsey 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 metaproteomics software are the development of proteomics tool libraries such as OpenMS (Sturm et al., 2008), Compomics (Barnes 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), (Algarn 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 (Mesureu 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
consequence, the estimated FDR and thus, the threshold for accepting protein identifications are higher and may lead to the rejection of true protein identifications.

Optimal databases would only include proteins and posttranslational modifications present in the sample and detectable by MS. However, taxonomic composition and protein abundance are usually unknown for environmental samples. Furthermore, protein content between analyzed samples may differ significantly. Therefore, database selection is a challenging task (Muth et al., 2015b; Tanca et al., 2016). This issue is further complicated by the adherence of the research community to the FDR concept (Muth et al., 2015b).

Originally Elias and Gygi, (2007) established the FDR concept for comparable protein identification in pure culture proteomics. In particular, the FDR enables comparability between different mass spectrometers and database search algorithms. Subsequently, the proteomics community accepted the FDR calculation as the standard to control the quality of protein identifications. An FDR of 1% was defined as threshold (Barnouin, 2011). However, a condition for the successful estimation of the FDR is that the database fits well to the sample. This is not guaranteed for metaproteomics studies, resulting in inaccurate approximations of the FDR. Therefore, it would be desirable that the metaproteomics community revises the FDR concept questioning the significance present in the sample and detectable by MS. Sometimes it is even possible to reconstruct the whole genome of single microorganisms within the microbial community, which is called binning.

Since these assembled metagenome protein databases match the actual sample, FDR estimation should be valid. However, the bioinformatic workflow to assemble metagenomes can also influence the protein identification (Tanca et al., 2016). For example, during metagenome assembly redundant reads where only one amino acid differs are sometimes condensed into a single read. This ignores protein isoforms and can lead to the loss of protein identifications. In contrast, a high number of translated reads in a database decrease protein identifications due to an increase in the FDR. In line with these problems, some authors experienced a higher number of protein identifications with read databases instead of contig databases (Timmins-Schiffman et al., 2017). Better protein identification was also observed by Tang et al. (2016) applying a graph-centric usage of reads as database.

The sequencing of metatranscriptomes is similar to metagenome sequencing [Fig. 2A]. In principle only translation of RNA to DNA is required. Identification of proteins against metatranscriptomes is beneficial, since organisms only transcript genes that are currently used (Wilmes et al., 2015). Sequencing a metagenome or metatranscriptome for each sample is not always possible due to the high cost and effort for the sequencing and the data processing. Thus, researchers use metagenomes from similar samples or comprehensive databases such as UniProtKB/SwissProt, UniProtKB/TrEMBL (UniProt, 2015), UniRef (Suzek et al., 2007), NCBI (Coordinators, 2017) or Ensemble (Yates et al., 2016) [Fig. 2A]. Database searches against complete comprehensive databases require long computation times and decrease the number of identified proteins due to the overestimation of the FDR. Reasonable constraints on these comprehensive databases are therefore necessary. For example Jagtap et al. (2013) proposed to search in two steps. Taxonomies or proteins identified in the first error-tolerant search are used to restrict the protein database for the second search [Fig. 2B]. This obviously increases computation time, but reduces the FDR and the threshold for protein identifications. In the end more proteins are identified, but how well this approximates the real FDR remains unclear. Another option for reduction of the FDR is to perform several searches against smaller sub databases and to merge their results afterwards (Muth et al., 2016; Tanca et al., 2016) [Fig. 2C]. A more reasonable approach to constrain the protein database is taxonomic foreknowledge, because in some cases taxonomic composition of the sample is known (Tanca et al., 2016) [Fig. 2D]. For example,

<table>
<thead>
<tr>
<th>Issue</th>
<th>Solution/bioinformatic strategy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grouping of redundant homologous proteins</td>
<td>1. Flexible grouping to metaproteins based on protein, peptide and taxonomy similarity</td>
<td>MetaProteomeAnalyzer (Muth et al., 2015a)</td>
</tr>
<tr>
<td>2. Grouping by shared peptide</td>
<td>ProPhae (Schneider et al., 2011)</td>
<td></td>
</tr>
<tr>
<td>Database tailoring</td>
<td>1. Two step database search</td>
<td>Jagtap et al. (2013)</td>
</tr>
<tr>
<td>2. Metapeptide database</td>
<td>May et al. (2016)</td>
<td></td>
</tr>
<tr>
<td>3. A “Graph-Centric Approach”</td>
<td>Graph2PeP/Graph2Prot (Zhang et al., 2016)</td>
<td></td>
</tr>
<tr>
<td>Taxonomic and functional evaluation</td>
<td>1. Calculate taxonomic value for each identified peptide (LCA) and visualize results</td>
<td>UniPept (Mesure et al., 2015)</td>
</tr>
<tr>
<td>2. Calculate taxonomic value for peptides using peptide similarity estimation and expression level weighting</td>
<td>Pipaslic (Penzlin et al., 2014)</td>
<td></td>
</tr>
<tr>
<td>3. Taxonomic evaluation (LCA) and functional prediction using RPSBLAST or HMMER3</td>
<td>ProPhae (Schneider et al., 2011)</td>
<td></td>
</tr>
<tr>
<td>4. Taxonomic (LCA) and functional evaluation using ECs, KEGG Ontologies and KEGG Pathways</td>
<td>Megan CE (Huson et al., 2016)</td>
<td></td>
</tr>
<tr>
<td>5. Taxonomic (LCA) and functional evaluation using UniProt Keywords, ECs, KEGG Ontologies, KEGG Pathways. Unknown sequences can be annotated using BLAST.</td>
<td>MPA (Muth et al., 2015a)</td>
<td></td>
</tr>
<tr>
<td>Storing and deployment of big data</td>
<td>1. Scalable set of sequence databases and specific database search algorithm</td>
<td>Compile and Blazmass (Chatterjee et al., 2016)</td>
</tr>
<tr>
<td>Quantitation</td>
<td>1. Detection and quantification of isotope ratios for Protein-SIP</td>
<td>MetaProSip (Sachsenberg et al., 2015)</td>
</tr>
</tbody>
</table>

Table 1
Overview about metaproteomic specific issues and appropriated software resp. bioinformatic strategies.
sequencing of the 16S-rRNA gene provides a taxonomic profile. Nevertheless, performing pre-searches against all taxonomies can help to avoid excessive constraints on protein taxonomy during the actual searches.

A smart idea to decrease computational time for protein database searches was recently proposed by May et al. (2016). They searched against peptide databases instead of protein databases [Fig. 2E]. This reduces the size of the search space due to the grouping of identical peptides from homologous proteins.

To summarize, all strategies to constrain protein databases carry some pitfalls and we would recommend researchers to try different approaches. Despite all these strategies for protein database construction, inaccurate FDR estimation hampers metaproteomics studies. Solutions other than the target-decoy approach are required to validate protein identifications across different MS and database search algorithms. A promising step towards this direction represent semi-supervised machine learning algorithms such as the software tools Percolator (Kall et al., 2007) or Nokoi (Gonnelli et al., 2015). They distinguish correct and incorrect peptide-to-spectrum matches using a classifier based on learning algorithms from real data.
5. Construction of user databases for protein identification: a use case

In order to visualize the impact of user databases a case study was conducted for a metaproteome analysis of three different biogas plant samples (BGP01, BGP02, BGP03). After phenol extraction, SDS-PAGE separation into ten fractions (Heyer et al., 2013) and LC–MS/MS measurement using an Orbitrap Elite (Heyer et al., 2016) different protein databases were tested [Fig. 3]. First the samples were searched against the UniProtKB/SwissProt database. Second several metagenomes from biogas plants were tested (metagenome 1, metagenome 2, metagenome 4, metagenome 5 (Stolze et al., 2016), metagenome 6 (Schlüter et al., 2008)). Of these metagenomes number 1 and 2 were prepared for BGP01 resp. BGP02. A metagenome from a waste water treatment plant (WWTP) (Püttker et al., 2015) was used as a negative control. Furthermore, the impacts of combining databases as well as of combining the results were evaluated.

The smallest numbers of identified metaproteins could be identified by the protein database search against the WWTP metagenome followed by the search against the UniProtKB/SwissProt database. Better results were obtained with the biogas plant metagenomes. Instead of 900 metaproteins for the protein database search against UniProtKB/SwissProt database about 2.000 metaproteins were identified using the biogas plant metagenomes. In some cases metagenomes appeared to be interchangeable, because metagenomes from other biogas plant samples showed equal or even better numbers of identified metaproteins as matching metagenomes, e.g. BGP02 and metagenome 2. This result questions whether the generation of a corresponding metagenome for each sample is always necessary. The combination of different metagenomes additionally increased the number of identified metaproteins to about 4,000 (combination metagenome 1+2+4+5+6). However, the number of additional metaprotein identifications decreased for each additional metagenome included in the search. In contrast the combination of metagenome 5 and the poorly matching metagenome from a waste water treatment plant (WWTP) decreased the number of identified metaproteins showing that an increased size of the database led to an increased chance of false positive hits and an increased FDR. The highest number of identified metaproteins was obtained with the separate search against all metagenomes (metagenome 1;2;4–6) and subsequent combination of the results. Focusing
on central metabolism and plotting the metaproteins into KEGG map 1200 clearly shows a higher coverage of pathways using the combined single searches (Fig. 4). This strategy avoided the increase of the FDR due to the bigger database, but the statistical correctness of this approach is questionable. However, it circumvents the accumulation of redundant sequence data in a combined database contributing to increased database size and FDR. Therefore, the removal of redundancy using peptide based databases could be a strategy to combine databases without increasing the FDR. Furthermore, the fact that combined metagenomes outcompete single corresponding metagenomes points out that many metagenome sequences do not comprehensively represent the microbial communities.

6. Protein inference problem and the grouping of proteins into “metaproteins”

Redundant identifications arising from homologous proteins share identical peptides and are therefore indistinguishable from each other. This hampers result evaluation and sample comparison within metaproteomic studies.

For pure culture proteomics Nesvizhskii et al., (2003) proposed to use the least number of proteins to explain all peptides. But this neglects...
the presence of protein isoforms or proteins from unsequenced microorganisms (Hettich et al., 2013) often found in analyses of metaproteomics data. To solve this issue the metaproteomics community started to develop concepts for grouping of redundant protein identifications [Table 2]. The metaprotein concept, introduced by Muth et al. (2015a), provides a good summary on protein grouping. Similar amino acid sequences (protein rules) or shared peptide identifications (peptide rules) constitute suitable criteria for grouping of homologous protein identifications into metaproteins. Conveniently, UniRef Clusters (Lu et al., 2014; Suzek et al., 2007) and KEGG Ontologies (Gotelli et al., 2012; Kanehisa et al., 2016) already classify most proteins on their sequence similarity. An easy retrieval of these classifications is enabled by the UniProtKB database, which is accessible through the UniProtJAPI library (Patient et al., 2008). Alternatively, proteins can be grouped when they share at least one identified peptide (Kohrs et al., 2014; Lu et al., 2014) or an identical peptide set (Keiblinger et al., 2012; Kolmeder et al., 2012; Schneider et al., 2011). It should be noted that for peptide comparison, the isobaric amino acids leucine and isoleucine are not distinguishable from each other.

All these strategies reduce the redundancy of the protein identifications successfully. However, only grouping based on identified peptides considers different conservation levels of the protein sequences. Thus, it enables a better taxonomic classification. Unfortunately, sample comparison using the peptide rule requires the protein grouping across all samples. Furthermore, the grouping may change as soon as additional samples are added. In consequence, grouping according to sequence similarity, such as UniRef clusters, is better suited for sample comparisons (Heyer et al., 2016; Kohrs et al., 2017).

In some instances it is desirable to consider the production of homologous proteins by different species. Homologous proteins often share peptides, which only differ in one or two amino acids. This indicates that these proteins should not be grouped together. To consider this bioinformatically, the Levenshtein distance (Levenshtein, 1966) between peptides of a protein group can be calculated (Muth et al., 2015a). Taxonomic foreknowledge is another option to improve metaprotein grouping. Protein groups can be restricted to certain phylogenetic affiliations, e.g. only proteins from the same genus.

### 7. Taxonomic and functional result evaluation

Comprehensive metaproteomics studies aim to describe the taxonomies and functions of complete microbial communities. In particular, the functions performed by each taxon should be elucidated. Protein taxonomy [Table 3] is usually defined according to the NCBI Taxonomy (Federhen, 2012). It comprises the classification for all taxonomic levels into the phylogenetic tree starting from species, genus and family via class, order and phylum to the kingdom and superkingdom levels.

In contrast to pure culture proteomics, a large portion of identified peptides in metaproteomics may belong to several proteins from different species. Thus, the taxonomic value of an identified peptide is estimated using the lowest common ancestor (LCA) of the protein taxonomies where this peptide occurs. Protein taxonomy is then defined as the LCA of the peptide identifications (Huson et al., 2011; Jagtap et al., 2012) or on the basis of unique peptides (Karlsson et al., 2012; Rooijers et al., 2011). Certain taxa have a much larger number of unique peptides, which biases the taxonomic profile towards these taxa. In general, unique peptides are fairly uncommon, as the analyses by UniPep demonstrate (Mesuere et al., 2015). The LCA approach is imprecise as well, because peptide taxonomy is often assigned on the order level and not on the species level. To refine the taxonomy profile Huson et al. (2016) propose to weigh the identified peptides and their LCA taxonomy by the amount of unique peptides. Another approach to improve the precision of the taxonomic profile is to weigh identified peptides by their spectral count and their occurrence in reference proteomes (Penzlin et al., 2014). Still, evaluation and comparison of taxonomic profiles is often challenging due to the high complexity of the data. This has led to several new approaches for data evaluation and visualization. The Krona plot (Ondov et al., 2011) clearly visualizes the taxonomy profile of a sample over all taxonomic levels. Furthermore, calculating community indices such as richness and evenness can give a general overview about the taxonomic profile of different samples (Heyer et al., 2016; Marzorati et al., 2008). In addition, specific interactions between single taxa can be examined by co-occurrence networks (Heyer et al., 2016; Huson et al., 2016; Jensen et al., 2001).

Several approaches with varying degree of specificity exist to assign functions to proteins [Table 3]. The protein acetyl-coenzyme A synthetase (P27550) is selected as example. It belongs to the acetate catabolism, which is sufficient to classify this proteins function. In other cases however, it is necessary to know that this protein transfers a coenzyme or contributes to chemotaxis. Originally, researchers studied the function of proteins separately through biochemical assays. Later their results were compiled, standardized and stored in databases. Recently, the functions of proteins from new species are derived from sequence similarity to functionally classified proteins. Functional classification of proteins with similar sequences is provided by databases such as KEGG ontology (KO) (Kanehisa et al., 2016), cluster of orthologous groups (COG) (Tatusov et al., 2000) and evolutionary genealogy of genes: non-supervised orthologous (eggNOG) (Huerta-Cepas et al., 2016).

Proteins of the same function possess differences in their amino acid sequence, but the sequences of their functional domains are highly conserved. Accordingly, the PFAM (Finn et al., 2016), the TIGRFAM database (Haft et al., 2013), the SMART database (Letunic et al., 2015) and the InterPro database (Finn et al., 2017) provide a functional classification based on similar functional domains. For example, acetyl-coenzyme A synthetase (P27550) possesses an acetyl-coenzyme A synthetase domain and an AMP-binding enzyme domain.

### Table 2

<table>
<thead>
<tr>
<th>Rule</th>
<th>Principle</th>
<th>Explanation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein rule</td>
<td>1. UniRef-Cluster</td>
<td>Grouping of proteins when they have 50%, 90% or 100% sequence similarity. Protein clustering provided by UniRef Cluster [Suzek2007].</td>
<td>Lu et al. (2014), Suzek et al. (2007)</td>
</tr>
<tr>
<td></td>
<td>2. KEGG Ontologies</td>
<td>Grouping of proteins when they are similar to functional classified genes within KEGG Ontology [Mai 2005]. KEGG Ontologies are provide by UniProtKB databases [JAPI PAPER].</td>
<td>Gotelli et al. (2012), Kanehisa et al. (2016)</td>
</tr>
<tr>
<td>Peptide rule</td>
<td>1. Shared peptide set</td>
<td>Group proteins when they share the same peptides.</td>
<td>Keiblinger et al. (2012), Kolmeder et al. (2012), Schneider et al. (2011)</td>
</tr>
<tr>
<td></td>
<td>2. One shared peptide</td>
<td>Group proteins when they have one identified peptide in common</td>
<td>Kohrs et al. (2014), Lu et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>3. One shared peptide + Levenshtein distance &lt; 2</td>
<td>Group proteins when they share the same peptides, but not if they have two similar peptides with less than 2 point mutations differences. This tracks the production of one protein by different microorganisms.</td>
<td>Muth et al. (2015a)</td>
</tr>
<tr>
<td>Taxonomy rule</td>
<td>1. Phylogenetic affiliation</td>
<td>Extends other rules by a certain phylogenetic affiliation.</td>
<td>Muth et al. (2015a)</td>
</tr>
</tbody>
</table>
Table 3
Strategies for taxonomic and functional annotation of proteins.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Name/principle</th>
<th>Explanation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxonomic classification</td>
<td>1. Lowest common ancestor</td>
<td>Define taxonomy as the lowest common ancestor into the phylogenetic tree.</td>
<td>Huson et al. (2011), Jagtap et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>2. Weighted lowest common ancestor</td>
<td>Adjust the lowest common ancestor by unique identification for the single taxa.</td>
<td>Huson et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>3. Peptide similarity estimation and expression level weighting</td>
<td>Weight taxonomy of identified peptides by their spectra abundance and their occurrence in a reference proteome.</td>
<td>Penzlin et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>4. Unique peptides</td>
<td>Define taxonomy and taxonomy profiles only based on unique peptides.</td>
<td>Rooijers et al. (2011), Karlsson et al. (2012)</td>
</tr>
<tr>
<td>Functional classification</td>
<td>1. KEGG Orthologies (KO)</td>
<td>Grouping of genes with same function by sequence similarity.</td>
<td>Kanehisa et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>2. Cluster of orthologues genes (COG)</td>
<td>Grouping of genes with same function by sequence similarity.</td>
<td>Tatusov et al. (2000)</td>
</tr>
<tr>
<td></td>
<td>4. Pfam</td>
<td>Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model).</td>
<td>Fins et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>5. TIGRFAM</td>
<td>Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model). In contrast to PFAM TIGRFAM emphasize protein function and enables a more precise functional classification.</td>
<td>Haft et al. (2013)</td>
</tr>
<tr>
<td></td>
<td>7. InterPro</td>
<td>Functional analyses of protein sequences by classifying them into families and predicting the presence of domains and important sites. Signatures are provided by 14 different member databases (among others PFAM, TIGRFAM, SMART).</td>
<td>Fins et al. (2017)</td>
</tr>
<tr>
<td>Pathway mapping</td>
<td>1. MetaCyc</td>
<td>Curated database of experimentally confirmed metabolic pathways.</td>
<td>Caspi et al. (2016)</td>
</tr>
<tr>
<td>Calculation of sequence similarity</td>
<td>1. BLAST</td>
<td>Calculation of sequence similarities.</td>
<td>Altschul et al. (1990)</td>
</tr>
<tr>
<td></td>
<td>2. DIAMOND</td>
<td>Calculation of sequence similarities. Up to 20,000 faster as BLAST.</td>
<td>Buchfink et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>3. MS-BLAST</td>
<td>Calculation of sequence similarities optimized for peptides.</td>
<td>Shevchenko et al. (2001)</td>
</tr>
</tbody>
</table>

It is important to note that functional annotation of proteins can be divided into categories such as molecular function, biological process or ligand, which are organized hierarchically. This is achieved by gene ontologies (GO) (Ashburner et al., 2000) and UniProtKB keywords (UniProt, 2015). For acetyl-coenzyme A synthetase (P27550) the UniProtKB keyword of the category ligand is ATP-binding protein; which belongs to the group of nucleotide-binding proteins. Enzyme commission numbers (EC) are another functional characterization of proteins (Bairoch; 2000). They use a four digit number code to classify enzymes depending on the catalyzed biochemical reaction. The EC for acetyl-coenzyme A synthetase (P27550) is 6.2.1.1; where 6 classifies it as a ligase; 6.2 as forming carbon sulfur bonds; 6.2.1. as acid-thiol ligase and 6.2.1.1. as acetate Co A ligase.

Conveniently, access to this taxonomic and functional metainformation is already provided by well annotated databases, such as UniProtKB. The entire database is available via the UniProt webpage and can be accessed programmatically via connectors such as the UniProtJAPI (Patent et al., 2008). Metagenomes miss taxonomic and functional annotation. Therefore, metagenome sequences are annotated by BLAST (Altschul et al., 1990) to link them to sequences of annotated proteins. Contigs may contain several genes with different functions, which can lead to false annotations. Moreover, the best BLAST hit is not always the correct one (Timmins-Schiffman et al., 2017) and for searches with short sequences, such as peptides, parameters for the BLAST should be adapted (MS-BLAST (Shevchenko et al., 2001)). Moreover, BLAST requires extensive computational time, which was addressed by development of the time-saving DIAMOND tool (Buchfink et al., 2015).

Another aim of metaproteomics studies is the analysis of certain metabolic pathways. Therefore, identified proteins can be visualized in the different metabolic and interaction pathways, using the pathway repositories MetaCyc (Caspi et al., 2016), KEGG pathways (Kanehisa et al., 2016) and Reactome (Fabregat et al., 2016). For KEGG pathways the web-based Interactive Pathways Explorer (iPath) (Yamada et al., 2011) provides an improved visualization and supports pathway analysis. Mapping of proteins to pathways is provided via the EC and KO numbers. Unfortunately, metabolic networks are incomplete, since many pathways are still unknown or specific for a minority of species. To overcome this limitation researcher started to create their own metabolic pathway maps. To achieve this, biochemical reactions, represented by EC numbers of identified proteins, were connected (Tobalina et al., 2015). A similar approach was chosen by Roume et al. (2015) aiming to identify key functions within a microbial community. Metabolic networks were modelled as a graph, where proteins (KO number) represented nodes and metabolites represented edges. Finally they defined key functions as nodes with high neighborhood connectivity. In future, networks based on metaproteome data could be used to predict metabolic fluxes, using software tools such as the CellNetAnalyzer (Klamt et al., 2007).

8. Quantitative data analysis in metaproteome studies

Protein quantification is crucial for comparative metaproteomics
studies. Indeed different approaches for quantitative proteomics exist, e.g. isotopic chemical labelling of peptides (Vauidel et al., 2010). But due to interference of these approaches with contaminating compounds many metaproteomics studies simply rely on the estimation of protein amount by counting identified peptides or spectra and normalizing these results (Ishihama et al., 2005; Zbylakov et al., 2007). Depending on data-dependent selection of precursor ions and successful peptide identification these approaches are inaccurate and possess a small dynamic range (Tabb2009). The quantification of the peptide peak intensity or area (Griffin et al., 2010) using tools such as Progenesis QI (http://www.nonlinear.com/progenesis/qi-for-proteomics/) or MaxQuant (Tyanova et al., 2016) is preferable. Alternatively, data-independent acquisition of MS/MS data (SWATH, MS²) combines peptide identification and quantification capturing all possible fragment information of all precursors for subsequent protein quantification from complex data (Bilbao et al., 2015). The most accurate quantification can be achieved by targeting only a single peptide (“single reaction monitoring”) or a limited selection of peptides of a certain protein (“single reaction monitoring”). For example, Saito et al. (2015) used this approach to quantify two nitrogen regulatory proteins for cyanobacterial taxa within microbial samples from the Central Pacific Ocean. The addition of isotopically labeled peptide for absolute quantification and the application of the Skyline software (MacLean et al., 2010) further improve this approach.

However, selection of peptides for targeted metaproteomics is more challenging than in pure culture proteomics, because a peptide may belong to multiple proteins from different taxa. Thus, the Unique Peptide Finder of the UniPept webservice (Mesure et al., 2016) was developed to facilitate the selection of unique peptides for a certain taxa.

9. Strategies for storing and deployment of huge data

Metaproteomics experiments comprise a massive amount of data including MS spectra, identified peptides and proteins as well as taxonomic and functional information. Our latest large-scale metaproteomics study produced about two Terabyte of data comprising roughly 15 million spectra and 23,000 identified metaproteins (data not shown). Consequently, appropriate data storage using a database management system (DBMS) is beneficial. Key challenges for DBMS are high speed for writing and reading data as well as efficient data storage. Since MS acquisition and search algorithms are relatively slow, writing speed has a negligible impact. In contrast, reading speed can be limiting, because researches want to evaluate all data at once. Furthermore, lists of thousands of proteins are unfeasible when inspecting results. Instead, researchers favor meaningful summaries, comparisons and intuitive visualizations. But this requires demanding database queries.

Relational database management systems, which use the “Structured Query Language” (SQL), have been the norm to manage data in the past. In recent years, alternatives to SQL have gained popularity and are aggregated under the term NoSQL (“Not only SQL”). Relational database management systems store data in separate tables, which are connected via unique relations. NoSQL database management systems use other concepts to store data like key-value associations (Berkeley DB (http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index.html, retrieved: 09-02-2017)), columns (Apache Cassandra (http://cassandra.apache.org/, retrieved: 09-02-2017)), documents (MongoDB (https://www.mongodb.com, retrieved: 09-02-2017)) or graphs (Neo4j, (www.neo4j.com, retrieved: 09-02-2017)).

NoSQL databases where motivated by the disadvantage in SQL databases to store all data in one place. In an analogy SQL databases can be imagined as a large building, which only a limited number of persons at a time can enter. An SQL query would be a person searching the building and collecting the information requested. If too many people search the building at a time, they will hinder each other and slow down the query process. NoSQL databases aim to address this issue of scalability. For instance, in our analogy Apache Cassandra creates a new identical building as soon as too many people try to enter. In consequence, NoSQL databases can handle more and more complex data requests. The disadvantage of NoSQL databases is reduced data consistency and large hard disc requirements due to multiple instances of the databases.

In sum NoSQL databases are highly beneficial for metaproteomics data. In line Chatterjee et al., (2016) used MongoDB for storing sequence information and Muth et al. (2015a) Neo4j for flexible result queries. Additionally, Mesure et al., (2015) are planning to use Berkeley databases to store the taxonomic value of each tryptic peptide.

Another trend of data storing and deployment which could be useful to increase the speed of data processing in metaproteomics is fast data (Braun et al., 2015). The fast data approach makes it possible to stream single spectra data to the cloud and process the data in real time for storing the results into the database. In other words, it parallelizes the data processing step and the measurement step to reduce experiment time. For example already the software MaxQuant Real-Time (Graumann et al., 2012) picks up this idea and processes the MS data in real time.

10. Future challenges, perspectives and demands

Predictions about the future of metaproteomics software need to anticipate future applications for metaproteomics. Foreseeable trends are an increase in MS resolution and therefore more data that will be acquired. Since metaproteomics is still an emerging field, an increase in the number of research studies about complex microbial communities is expected. A great potential for the application of metaproteomics are process control in technical applications as well routine diagnostics of fecal samples. So far it is known that microbial communities in the human gut system are linked with autoimmune and allergic diseases, obesity, inflammatory bowel disease (IBD), and diabetes (Clemente et al., 2012). Consequently, the number of samples in clinical settings could rise to several thousand per day. Such an increase in sample numbers requires software tools that can handle huge data amounts. For routine diagnostics the total computation time may not exceed a few hours, so that a complete metaproteomics analysis may require less than one day. Another aspect is that software for medical applications has to conform to high quality standards and specific privacy regulations. Moreover, medical staff without a special bioinformatic background should be able to operate such software tools. Although the routine usage of metaproteomics is still in question, the development may proceed quickly. For example, MALDI-MS based identification of microbial isolates became a standard procedure in clinical laboratories.

Strategies to facilitate software usage are to provide it via Docker (e.g. Bioconda https://bioconda.github.io/, retrieved: 09-02-2017) or web services to avoid problems with the installation and configuration of complex software frameworks. For example, developers of the MPA are planning to provide their software platform as web service within the de.NBI project. Most users with a medical or biological background would favor a graphical ready-to-use software tool. In contrast, bioinformaticians prefer modular software packages operated from the command line. The latter strategy enables flexible assembly of workflows and an easy improvement of single modules. The challenge for future development of metaproteomics software is to satisfy both sides.

Because metaproteomics is still a developing field, universal standards still have to be adopted by the community. Implementation of ring trials for metaproteomics data processing could further insights into the comparability of software tools, and enable the introduction of quality standards.

Further improvement requires the validation of protein identifications by the FDR estimation. In contrast to pure culture proteomics the estimated FDR is not always correct since the protein sequences for the
investigated samples are often unknown. A solution might be the usage of semi-supervised machine learning algorithms such as the software tools Percolator or Nokoi (Gonnelli et al., 2015).

The use of protein databases could be standardized as well. While some researchers use comprehensive protein databases, others use diverse metagenomes, which differ in the processing state and origin. A solution might be the generation of non-redundant (May et al., 2016), fusion metagenomes for each type of microbial community. Thereby, this fusion metagenome should be assembled as far as possible.

Additionally, the binning of metagenomes may also improve the protein database quality. Proteins of the same function or metabolic pathway are often located adjacent on a contig or operon. Thus, they should feature equal expression patterns.

The key to handle the increased amount of data is the real-time processing of all arising MS data as well as the scalability of the software and the database. This means that the single computational steps operate in parallel and hardware resources can be allocated on demand, e.g. by cloud computing (Mell and Grance, 2010). To guarantee the long term maintenance and support for such systems, it is reasonable to follow the latest trends from the industry instead of developing own solutions. Suitable frameworks, among others, are Apache Spark (http://spark.apache.org/, retrieved: 09-02-2017) for analyzing data distributed in the cloud and OpenStack (https://www.openstack.org/, retrieved: 09-02-2017) to manage the instances running on the cloud.

Another strategy to decrease computation time is the smart deployment of hardware resources. Graphical processing units (GPU) can perform specific calculations in parallel. On the other hand central processing units (CPU) are suited for general tasks, but work serially. Identification of MS/MS spectra is a calculation that can be parallelized. In line, the protein database search algorithm X!Tandem was recently adopted to utilize a GPU (He and Li, 2015).

Beside adaptation of metaproteomics to bigger data volumes and the decrease of computation time, improved bioinformatic strategies are required to increase the number of identified spectra. State-of-the-art metaproteomics studies only achieve identification of 5–30% spectra. An estimated 30% of all spectra belong to solvent and background components (Griss et al., 2016). This means at least another 30% spectra remain unidentified. Better metaproteomics software should contribute to overcome this issue. The generation of more suitable metagenomes for protein identification may increase the amount of identified spectra significantly. Inversely, assembly of metagenomes can be validated using peptides identified in metaproteomics studies (Nesvizhskii, 2014). Spectral libraries represent another strategy to handle unidentified spectra (Lam et al., 2007). They could store and cluster spectra from any sample. Samples can be also compared based on their unidentified spectra. Interesting spectra can be annotated later using protein database search algorithms. Due to the drastic reduction of candidates, manual de novo sequencing is also possible (Frank and Pevzner, 2005). Function and taxonomy of de novo peptides can be derived by MS-BLAST search (Shevchenko et al., 2001). However, de novo sequencing of peptides is hampered by the short length of tryptic peptides which impede MS-BLAST identification. Better de novo and MS-BLAST results could be achieved by other proteases such as Lys-C (Jekel et al., 1983) or Arg-C, which result in longer peptides. Due to increased computational power and more precise MS it may become possible to search against a database containing all theoretical peptides for a specific mass (Sadygov, 2015). This would also solve problem with the database size dependency of the FDR estimation.

Finally, metaproteomics software can benefit from the incorporation of data from other multi-omics techniques (Brink et al., 2016; Heinitz-Buschart et al., 2016), e.g. metabolome data. For a detailed overview on multi-omics data processing, please refer to Franzosa et al. (2015).

11. Conclusions

Metaproteomics represents a powerful tool for the taxonomic and functional characterization of complex microbial communities from environmental samples. In the future it has the potential to become a valuable tool for routine diagnostics, e.g. analysis of human feces. However, success of metaproteomics studies depends on dedicated software tools. These tools must be capable to handle big data, but also need to be usable by people with no background in bioinformatics. To achieve these goals, web services and software tools capable of parallel computing are reasonable (e.g. cloud computing). This would decrease computational costs and enables small laboratories to perform metaproteomics studies. Moreover, metaproteomics studies will benefit from software supporting the taxonomic and functional interpretation of results. Even if it is obvious, the close cooperation of bioinformaticians and biologists should also be considered during software development.

Competing interests

The authors declare that they have no competing interest.

Funding

This work was supported by the Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) communicated by the Agency for Renewable Resources (FNR), grant no. 22404115 (“Biogas Measurement Program III”) and the de.NBI network (“MetaProtServ de-NBI-039”).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author’s contributions

The manuscript was written by Robert Heyer (RH), Dirk Benndorf (DB), Kay Schallert (KS), Beatrice Becher (BB), Udo Reichl (UR) and Günther Saake (GS). All authors read and approved the final manuscript.

Additional files

Not applicable.

Availability of data and material

Not applicable.

Acknowledgement

Not applicable.

References

characterization of microbe samples using shotgun metagenomic sequencing. J. Proteome Res. 15, 2697–2705.


