The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation

Thilo Muth,†‡‡ Alexander Behne,‡‡ Robert Heyer,‡ Fabian Kohrs,‡ Dirk Benndorf,‡ Marcus Hoffmann,† Miro Lehteva,§¶ Lennart Martens,§¶ and Erdmann Rapp*†‡

†Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany
‡Chair of Bioprocess Engineering, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany
§Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium
¶Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium

Supporting Information

ABSTRACT: The enormous challenges of mass spectrometry-based metaproteomics are primarily related to the analysis and interpretation of the acquired data. This includes reliable identification of mass spectra and the meaningful integration of taxonomic and functional meta-information from samples containing hundreds of unknown species. To ease these difficulties, we developed a dedicated software suite, the MetaProteomeAnalyzer, an intuitive open-source tool for metaproteomics data analysis and interpretation, which includes multiple search engines and the feature to decrease data redundancy by grouping protein hits to so-called meta-proteins. We also designed a graph database back-end for the MetaProteomeAnalyzer to allow seamless analysis of results. The functionality of the MetaProteomeAnalyzer is demonstrated using a sample of a microbial community taken from a biogas plant.

KEYWORDS: bioinformatics, environmental proteomics, mass spectrometry, metaproteomics, microbial communities, software

INTRODUCTION

Mass spectrometry (MS)-based analysis of pure culture proteomes or simple mixed cultures has advanced rapidly in the past decade. There is now a growing interest in studying complex multispecies samples such as entire microbial communities. Microbial consortia are key players in geo-chemical cycles,1,2 biochemical networks, and biotechnological and medical applications.3–7 For instance, metaproteomics or whole community proteomics8 investigating enzymatic capabilities of microbes on the protein level is applied for wastewater treatment10 and biogas plants11 or in metaproteomic studies with clinical background by analyzing microbial proteins within the human oral cavity12 and intestinal tract.13,14 Over the past few years, the development of metaproteomics is driven mainly by the ability to sequence microbial genomes and metagenomes via high-throughput sequencing15 as well as by the simultaneous improvements in MS instrumentation that allow rapid high-resolution analysis of microbial community samples, predominantly by using shotgun proteomic approaches.16 Despite these improvements, however, the field still suffers from the exceptional complexity and heterogeneity of those samples, which hamper data evaluation.17 State-of-the-art protein identification algorithms,18,19 for instance, are designed to handle single-species samples and are severely challenged by size and redundancy of multispecies protein sequence databases.20 In addition, protein hits are typically returned from several hundreds of different species, which further exacerbates the already confounding protein inference problem.21 Taxonomic binning of the identified peptides therefore must be addressed in a sophisticated manner as well.22 Another challenging issue is encountered in the functional annotation of proteins, as metaproteomics research is not only interested in (single-)protein identifications, but also focuses strongly on specific functions performed by microorganisms in an ecosystem.23 Unfortunately, no stand-alone software tool currently exists to aid metaproteomics research in reliably addressing the central question of the field: “who is doing what?”. Here we describe the MetaProteomeAnalyzer (MPA), a free, open-source, end user oriented Java software suite for the comprehensive analysis and visualization of metaproteomics data sets (http://meta-proteome-analyzer.googlecode.com; a guided, hands-on tutorial can be found in Note 1, Supporting Information). For exhaustive peptide and protein identification, the MPA features four different freely available database search algorithms and furthermore allows for the integration of results derived from the commercial MASCOT search engine (version 2.4).18 The combination of these complementary search engines leads to an increase in protein and peptide identifications as well as in identification reliability. The MPA provides an intuitive workflow for the automated functional and taxonomic characterization of proteins of interest. The software allows for grouping of redundant proteins in the result set according to a set of provided rules, and it also offers an innovative way of querying the results of a metaproteomics...
analysis by providing a graph database-based back end. The MPA is designed as a client–server application, with the identification workload handled by a high-performance server. In addition, the local client provides a user-friendly graphical user interface to analyze and interpret the results. An example data set derived from a sample of a biogas plant (Note 2, Supporting Information) is included in the MPA viewer application, a stand-alone client version with limited processing capabilities.

**MATERIAL AND METHODS**

**General Workflow**

The MPA represents a software pipeline for the analysis and interpretation of metaproteomics data sets. An overview of the general workflow employed in the MPA is outlined in Figure 1 and is described in detail here. At the start of a new project (1), experimental data are provided, and the corresponding tandem mass spectra are loaded and sent to the processing server. The server then executes up to four different database search algorithms (X!Tandem, OMSSA, Crux, and InsPect) for peptide and protein identification and can also retrieve identifications from the widely used commercial Mascot software using mascotdatatfile. Search results from the different search engines are merged after their individual scores are converted to uniform significance measures, so-called q-values reflecting the minimum FDR for the identifications. The obtained identifications are complemented subsequently with additional information such as protein taxonomy and function (2). The MPA server also takes care of storing spectra, results, and annotations in a relational database and acts as a simple laboratory information management system (LIMS) that stores all data for later reanalysis or meta-analysis (see Note 3, Supporting Information for the database schema). The user can then retrieve the processed results from the server for performing analyses via the client application (3). In this step, proteins are grouped into so-called meta-proteins (see Note 4, Supporting Information for grouping rules) and are annotated with additional information derived from external resources: general protein-level information (e.g., ontology keywords) from UniProt, taxonomic information from NCBI, enzyme information from Enzyme Commission (E.C.) classification scheme, and metabolic pathway information from KEGG. It is also possible to add customized protein databases to the MPA workflow, e.g., for including protein sequences derived from metagenomic sequencing. The formatting and indexing of such user-defined FASTA databases are explained in more detail on the MPA wikipages.

Figure 1. MetaProteomeAnalyzer workflow. (1) Project and experiment management is handled on workflow start. (2) MPA server performs database searches and querying of meta-information. (3) Proteins are annotated and grouped to meta-proteins. (4) MPA client user interface provides a results overview with heatmap and charts. (5) Detailed top-down protein result views display meta-proteins and label-free quantification measures. (6) Results linked to meta-information on pathways, ontologies, enzymes, and taxonomies are displayed in additional views. (7) Results can be queried with user-defined functions on the graph database.
peptide across all search engines, and the fully interactive, annotated spectrum view for each peptide-to-spectrum match.34,35 Moreover, several label-free quantitative measures are provided for each protein. This includes spectral count, normalized spectral abundance factor (NSAF),36 and exponentially modified protein abundance index (emPAI).37 The enzyme and pathway views display proteins aggregated by E.C. numbers32 and KEGG pathways,33 which enables a direct inspection of microbial functions (6; Figure 3, Supporting Information). Further detailed views provide protein groupings based on meta-proteins, taxonomies, and ontologies. To focus on the possible role of a defined microbial group in the sample data set, flexible filtering methods are available in all views. The enzyme and pathway views display proteins aggregated by E.C. numbers32 and KEGG pathways,33 which enables a direct inspection of microbial functions (6; Figure 3, Supporting Information). Further detailed views provide protein groupings based on meta-proteins, taxonomies, and ontologies. To focus on the possible role of a defined microbial group in the sample data set, flexible filtering methods are available in all views. Perhaps the most innovative feature of the MPA is the Neo4j-based graph database (7) that holds all the protein meta-information, which allows the user to handle complex queries efficiently using a query dialogue (Figure 4, Supporting Information). For the example data set from a multispecies sample of a microbial community of a biogas plant, this feature of the MPA allows searching of specifically those species that are involved in the production of methane using the key enzyme Methyl coenzyme M Reductase. The power of this innovative approach is illustrated in more detail in Note 5 of the Supporting Information. Each rule can be applied individually or in combination with other rules, which will yield different results. To visualize the impact of meta-protein generation on the taxonomy, we used the Krona display.42

**Meta-Protein Generation**

In metaproteomics, identified peptides frequently belong to homologous proteins expressed by organisms belonging to different species, which causes redundant protein identifications to be reported.58 Previous studies propose several strategies to handle redundant protein hits.39-41 The MPA software incorporates several such approaches within its result processing workflow, the most fundamental being the grouping of proteins according to peptide similarity (Peptide Rule) and protein similarity (Protein Cluster Rule). To minimize the shortcomings of established methods, these rules can be further extended by taking the protein taxonomy into account (Taxonomy Rule). As a precondition for the taxonomy rule, information about the taxonomic lineage of proteins must be inferred. Therefore, the taxonomical process, which also affects peptides in relation to proteins and proteins in relation to meta-proteins, is embedded into the meta-protein generation workflow. The common ancestor rule determines the taxonomic lowest common ancestor by all shared peptides for the proteins. Conversely, the most specific taxonomy rule preserves peptide-level specificity for proteins, that is, usually the taxonomy on the species or subspecies level is conserved. As a starting point, a preliminary meta-protein is generated for every protein hit in the raw results. Subsequently, meta-proteins are fused by applying these rules. More detailed information on the meta-protein generation rules can be found in Note 4 of the Supporting Information. Each rule can be applied individually or in combination with other rules, which will yield different results. To visualize the impact of meta-protein generation on the taxonomy, we used the Krona display.42

---

**Figure 2.** Search result panel of MetaProteomeAnalyzer. The identified proteins are displayed in the top panel, the identified peptides for the currently selected protein in the middle-left panel, and the peptide-to-spectrum matches for the selected peptide across all search engines in the lower-left panel. The right panel shows the annotated fragment ion series of the currently selected peptide-to-spectrum match.
Graph Database System

In general, users have challenging questions during data investigation, and software solutions cannot predict all such use cases. To provide maximal flexibility and power in interrogating the metaproteomics results, the open source graph database Neo4j (www.neo4j.com) has been integrated into the MPA, which enables user-defined querying of the results based on the Cypher query language (http://docs.neo4j.org/refcard/1.9/). Instead of taking the classical approach of a relational database system with tables and indices, this database structure is modeled as a graph consisting of nodes (vertices) and relationships (edges). Both of these entities are named, and relationships are directed referring to a start and end node. Additionally, the graph database uses properties, which are basically key-value pairs that represent certain attributes for nodes and relationships. The graph database is a fully transactional database management system with Create, Read, Update, and Delete (CRUD) methods that are common to relational databases. Although the core of Neo4j has been developed in Java, it provides access to various application programming interfaces (APIs). In contrast to common databases, Neo4j offers two database modes as it runs in either embedded or server mode. The embedded version is used in this case to incorporate the database directly in the MPA client application. Furthermore, the embedded mode comes with the advantages of low latency and full control of the database life cycle. In the case of proteomics data, the graph database structure consists of representative node variants (Table 1) and relationships (Table 2).

### Table 1. Node Types and Descriptions for the Graph Database

<table>
<thead>
<tr>
<th>node type</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>Identified proteins; properties are protein accession, description, sequence coverage, species, and spectral count.</td>
</tr>
<tr>
<td>Peptides</td>
<td>Identified peptides; properties are peptide sequence and spectral count.</td>
</tr>
<tr>
<td>PSMs</td>
<td>Peptide−spectrum matches; properties are spectrum identifier and search engine score.</td>
</tr>
<tr>
<td>Taxonomies</td>
<td>Taxonomies; properties are taxonomy name, NCBI taxonomy ID, and rank.</td>
</tr>
<tr>
<td>Ontologies</td>
<td>UniProt ontologies; properties are ontology name and category (e.g., biological process)</td>
</tr>
<tr>
<td>Pathways</td>
<td>KEGG pathways; properties are KO number and KEGG description.</td>
</tr>
<tr>
<td>Enzymes</td>
<td>E.C.-based enzymes; properties are E.C. number and description.</td>
</tr>
</tbody>
</table>

### Table 2. Relationship Types and Descriptions for the Graph Database

<table>
<thead>
<tr>
<th>relationship type</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAS_PEPTIDE</td>
<td>$N_{out}$ Proteins; $N_{out}$ Peptides; relationship for proteins that share the peptides.</td>
</tr>
<tr>
<td>IS_MATCH_IN</td>
<td>$N_{out}$ PSMs; $N_{out}$ Peptides; relationship for PSMs that match for peptides.</td>
</tr>
<tr>
<td>BELONGS_TO</td>
<td>$N_{out}$ Proteins; $N_{out}$ Taxonomies; relationship for proteins that belong to certain taxonomies.</td>
</tr>
<tr>
<td>BELONGS_TO_ENZYME</td>
<td>$N_{out}$ Proteins; $N_{out}$ Enzymes; relationship for proteins that fulfill an enzymatic function.</td>
</tr>
<tr>
<td>BELONGS_TO_PATHWAY</td>
<td>$N_{out}$ Proteins; $N_{out}$ Pathways; relationship for proteins that are part of certain pathways.</td>
</tr>
<tr>
<td>INVOLVED_IN_BIPROCESS</td>
<td>$N_{out}$ Proteins; $N_{out}$ Ontologies (Biological Process); relationship for proteins that are involved in biological processes.</td>
</tr>
<tr>
<td>HAS_MOLECULAR_FUNCTION</td>
<td>$N_{out}$ Proteins; $N_{out}$ Ontologies (Molecular Function); relationship for proteins that have molecular functions.</td>
</tr>
<tr>
<td>BELONGS_TO_CELL_COMP</td>
<td>$N_{out}$ Proteins; $N_{out}$ Ontologies (Cellular Component); relationship for proteins that belong to cellular components.</td>
</tr>
<tr>
<td>IS_SUPERGROUP_OF</td>
<td>$N_{out}$ Enzymes; $N_{out}$ Enzymes; relationship to reflect the enzyme (E.C.) hierarchy.</td>
</tr>
<tr>
<td>IS_ANCESTOR_OF</td>
<td>$N_{out}$ Taxonomies; $N_{out}$ Taxonomies; relationship for the taxonomic hierarchy (from superkingdom to species).</td>
</tr>
<tr>
<td>IS_METAPROTEIN_OF</td>
<td>$N_{out}$ Proteins; $N_{out}$ Proteins; relationship between a meta-protein and a protein (see Note 5, Supporting Information).</td>
</tr>
</tbody>
</table>

### Experimental Data

The example data set represents the metaproteome of a complex microbial community derived from an agricultural biogas plant located in Magdeburg/Ebendorf (Saxony-Anhalt, Germany) and was obtained by liquid chromatography tandem mass spectrometry (LC−MS/MS). Main process parameters as well as substrate feed composition are summarized in Table 1 of the Supporting Information. More details on sample preparation, LC−MS/MS measurement, and data processing can be found in Note 2 of the Supporting Information.

### Database Searching

We performed the database searching using the search algorithms MASCOT (version 2.3),18 X!Tandem (version 2013.02.01),24 and OMSSA (version 2.1.8).43 MS/MS spectra were searched against UniProt/SwissProt database (version 2013/02/20). Trypsin was used as default enzyme cleavage parameter, and the maximum allowed number of missed cleavages was set to one. Carbamidomethylation of cysteine (Cys+57 Da) was chosen as fixed modification and oxidation of methionine (Met+16 Da) as variable modification. The precursor ion tolerance was set to 10 ppm, and the fragment ion tolerance was set to 0.5 Da. Target-decoy searching was performed, and the decoy database was constructed by reversing the protein sequences from the target database.

### RESULTS AND DISCUSSION

To evaluate the MPA and its processing steps, we conducted an experiment with a real data set taken from a biogas plant sample. First, we show the impact of using multiple search engines on the number of identifications. Then, we demonstrate the grouping of redundant proteins to meta-proteins on the exemplary data set. Finally, we illustrate the possibility of asking user-defined questions regarding specific aspects of the given data.

### Search Engine Comparison

To test the impact of using multiple search engines, the biogas plant data set was searched with three database search algorithms: X!Tandem, OMSSA, and MASCOT. In the following, we limited the comparison between the different search engines to the number of identified spectra and distinct peptides, as the number of reported proteins varied significantly between the search algorithms. By this restriction, we could guarantee a fair comparison between the different search engines. In the first step, we compared the number of identified spectra at 5% FDR (Figure 5a, Supporting Information). In the...
Euler diagram, it can be found that each search engine yielded a significant amount of unique spectrum identifications. X! Tandem provided the highest number of unique identifications on the spectrum level, that is, 799 (24.2%) out of 3295 identified spectra.

The next comparison involved the number of distinct peptides identified from each of the algorithms (Figure 5b, Supporting Information). The Euler diagram showed that X! Tandem again provided the highest number of peptide hits: 281 (28.3%) out of 992 peptides were identified. The impact of applying different meta-protein rules on the calculated composition of the microbial community was visualized by using the Krona display presenting attributed spectral counts onto the different taxonomic levels. In comparison to unprocessed data (Figure 11a, Supporting Information), data from meta-proteins sharing at least one peptide, which had been subsequently assigned to the common ancestor (Figure 11b, Supporting Information), showed a different taxonomic composition: a slightly higher proportion of Methanocarcinales (19% instead of 18%) and a lower proportion of the order Methanomicrobiales (1% instead of 8%) were found due to more assignments to the common ancestor phylum of Euryarchaeota (10% instead of 1%). The applied meta-protein rule (sharing at least one peptide, common ancestor) grouped proteins from wider phylogenetic ranges into meta-proteins and partially prevented the phylogenetic assignment on the order level. Less stringent meta-protein rules, for example, UniRef90 clustering (Figure 11c, Supporting Information), will allow more detailed assignment but also result in a smaller decrease of sample complexity. More differences in the microbial composition can be found when applying other protein grouping rules, such as UniRef50 (Figure 11d, Supporting Information) clustering and shared peptide sets (Figure 11e, Supporting Information).

The comparison of phylogenetic assignment of meta-proteins in the Krona displays is based on spectral counts. Unfortunately, the calculation of label-free quantification methods, such as NSAF36 or emPAI37 requires either a defined sequence length for the meta-protein or its complete amino acid sequence. The calculation of both measures is therefore biased by partial sequences of proteins in the databases. When working with metaproteome data, this problem becomes worse as the measures are based on the protein hit derived from the database: this protein sequence may differ from the actual protein in the sample. Therefore, we used the metric of spectral counts and added so-called aggregate functions (see Note 1, Supporting Information), for example, to calculate the average of the protein NSAF values for each meta-protein within a sample. This method can be used as a straightforward strategy to enable the label-free quantification for meta-proteins. Another possibility would be the application of an algorithm that is able to handle the high amount of shared peptides between homologous proteins from different organisms in a reasonable manner: for example, a tool called Pipasic (peptide intensity-weighted proteome abundance similarity correction) that corrects strain-level identification and quantification results based on spectral counts.36

Meta-Protein Generation

In this section, we investigate the results of the meta-protein generation. For this purpose, we used a group of proteins denoted as F420-dependent methylenetetrahydromethanopterin dehydrogenase (henceforth F420-MTHMO). The example data set features six such protein identifications. The initial state of ungrouped, redundant protein results is shown in Figure 6 of the Supporting Information. Each of the protein identifications was assigned to a single meta-protein at this stage. One way to reduce the redundancy of the results is to group the proteins according to common, so-called shared peptides. Because each of the proteins holds at least one peptide that is also associated with another protein, all proteins are grouped under a single meta-protein according to this grouping rule (Figure 7, Supporting Information). Note that the isomeric amino acids leucine (L) and isoleucine (I) are considered identical. However, as homologous proteins in nature often differ in their amino acid sequence, this exact string matching approach may lead to an incorrect assignment of possible candidates for the protein grouping. Consequently, we also incorporated a distance measure for peptides by allowing a defined maximum number of point mutations. A less stringent peptide similarity definition of one amino acid substitution results in more common peptide sets (Figure 8, Supporting Information). In contrast to the previous rule, the presence of a peptide unique to one specific protein (e.g., A3CSZ5) precludes this protein from also being grouped under the same meta-protein. This can actually be a desired outcome as unique peptides may be an indicator for the presence of distinct proteins.

Another rule presents the grouping of proteins according to UniRef database cluster assignments.44 When applying the UniRef50 rule to our example, the majority of all proteins fall into the same similarity cluster due to whole protein sequence similarities, apart from the protein O29544 that is derived from a distant species (Figure 9, Supporting Information).

The last rule, which we applied for protein grouping, uses the minimum of one shared peptide. In addition, however, we refined the results by using a taxonomic cutoff. This prevents grouping of proteins whose lineages converge above a specified taxonomic rank threshold. In this example, on the basis of the inferred protein taxonomy rule, the majority of all proteins is grouped under a single meta-protein (Figure 10, Supporting Information). In both cases, the protein O29544 is assigned to a discrete meta-protein belonging to the taxonomic class of Archaeoglobi. The F420-MTHMO is part of the incomplete methanogenesis of the thermophile sulfate reducing Archaeoglobus fulgidus.45 In contrast to methanogens, the pathway is meant to be used in reverse direction here: thus, oxidizing methyl groups derived from the acetyl-CoA decarboxylase synthase to carbon dioxide. The rules for meta-protein assignment thus correctly separate proteins with similar functional annotation according to their phylogenetic distance as well as to their actual activity.

As illustrated in this section, the choice of meta-protein generation rules has a profound effect on the number of the resulting meta-proteins. On the whole example data set, a reduction between 44 and 50% of the unprocessed data set could be achieved depending on the rules chosen (Table 2, Supporting Information). The impact of applying different meta-protein rules on the calculated composition of the microbial community was visualized by using the Krona display presenting attributed spectral counts onto the different taxonomic levels. In comparison to unprocessed data (Figure 11a, Supporting Information), data from meta-proteins sharing at least one peptide, which had been subsequently assigned to the common ancestor (Figure 11b, Supporting Information), showed a different taxonomic composition: a slightly higher proportion of Methanocarcinales (19% instead of 18%) and a lower proportion of the order Methanomicrobiales (1% instead of 8%) were found due to more assignments to the common ancestor phylum of Euryarchaeota (10% instead of 1%). The applied meta-protein rule (sharing at least one peptide, common ancestor) grouped proteins from wider phylogenetic ranges into meta-proteins and partially prevented the phylogenetic assignment on the order level. Less stringent meta-protein rules, for example, UniRef90 clustering (Figure 11c, Supporting Information), will allow more detailed assignment but also result in a smaller decrease of sample complexity. More differences in the microbial composition can be found when applying other protein grouping rules, such as UniRef50 (Figure 11d, Supporting Information) clustering and shared peptide sets (Figure 11e, Supporting Information).

The comparison of phylogenetic assignment of meta-proteins in the Krona displays is based on spectral counts. Unfortunately, the calculation of label-free quantification methods, such as NSAF36 or emPAI37 requires either a defined sequence length for the meta-protein or its complete amino acid sequence. The calculation of both measures is therefore biased by partial sequences of proteins in the databases. When working with metaproteome data, this problem becomes worse as the measures are based on the protein hit derived from the database: this protein sequence may differ from the actual protein in the sample. Therefore, we used the metric of spectral counts and added so-called aggregate functions (see Note 1, Supporting Information), for example, to calculate the average of the protein NSAF values for each meta-protein within a sample. This method can be used as a straightforward strategy to enable the label-free quantification for meta-proteins. Another possibility would be the application of an algorithm that is able to handle the high amount of shared peptides between homologous proteins from different organisms in a reasonable manner: for example, a tool called Pipasic (peptide intensity-weighted proteome abundance similarity correction) that corrects strain-level identification and quantification results based on spectral counts.36

Integration of Meta-Information from External Resources

Besides the provision of taxonomic information, the identified proteins also point to biological functions. To provide access to
such external meta-information, all protein hits are linked to external sequence databases (UniProt, NCBI), domain/family databases (Pfam, Interpro), services for sequence analysis (Protein BLAST), and functional annotation databases (KEGG, QuickGO). By using the pathway view in the MPA, all proteins identified in the data set can easily be plotted onto the respective KEGG pathway maps, displayed in Figure 3 as an example for the major pathway of carbon metabolism. By restricting the taxonomic range to the superkingdom, the major pathways can be assigned to Archaea (Figure 12, Supporting Information) and Bacteria (Figure 13, Supporting Information). The pathway of methanogenesis is, for example, exclusively present in Archaea and glycolysis/gluconeogenesis mainly in Bacteria. Another example is provided by the analysis of the amino acid synthesis pathway, which revealed that particular bacterial taxa show a preference for the synthesis of certain amino acids: Proteobacteria prefer arginine and lysine (Figure 14, Supporting Information), while Firmicutes favor tryptophan and histidine (Figure 15, Supporting Information). However, enzymes for amino acid synthesis are generally underrepresented in Archaea (Figure 16, Supporting Information). The high abundance of enzymes involved in arginine, lysine, tryptophan, and histidine synthesis in the metaproteome might indicate the de novo synthesis of these amino acids. By surveying the amino acid composition of maize bulk protein, these amino acids are underrepresented in relation to the abundance in microbial biomass: this supports the hypothesis that de novo synthesis is required for microbial growth in

Figure 3. KEGG pathway display. The major pathway of the carbon metabolism is shown. KEGG pathway map (1200) of the general carbon metabolism is shown as a net of metabolites and its intermediates. Edges connecting single metabolites by arrows represent respective enzymes and the potential direction of enzymatic conversion. Proteins identified in the data set are highlighted in red after submission from the MPA to the KEGG Web site, thereby representing the coverage of this pathway.
biogas plants. Both examples show that the MPA facilitates the integration and interpretation of such taxonomic and functional data.

Graph Database Driven Query System

In this section, we illustrate the use of querying the data via the flexible graph database system: it allows the user to address specific questions that are not straightforward to answer by the classical result views. To demonstrate the benefits of this approach, we performed four different queries of increasing complexity on the exemplary data set. The syntax and visual representation of these queries are described in Note 5 of the Supporting Information, and the query results are shown in Table 3 of the Supporting Information. For each of the following examples, the single shared Peptide Rule was taken for the protein redundancy reduction. The first query aims to retrieve all meta-proteins with their related proteins and peptides. This query starts at the protein nodes with the relationship IS_METAPROTEIN_OF and traverses the graph via proteins to peptides. The query resulted in 712 meta-proteins linked to 1351 constituent proteins identified by 1140 peptides. The second query extends the first query by a WHERE condition and a regular expression to exclude the term "keratin". Filtering out contaminant proteins such as keratin or searching for a specific protein is an important step in any proteomics analysis. This restricted query resulted in 702 meta-proteins linked to 1236 proteins and 947 peptides. To find out which taxa and functions can be associated with the data set, the identified meta-proteins are grouped by their taxonomies and ontologies (in this case the biological processes) in the next query: this resulted in 229 taxonomies (all levels), 150 ontologies (biological processes), 428 meta-proteins, and 820 proteins. The last query contains two MATCH clauses and the WHERE condition for the pathway identifier Methyl CoM Reductase (K00399), the final enzyme of methane production. In this case, the query returns the single protein P07962 (Methyl CoM Reductase Subunit Alpha) from the organism Methanosarcina barkeri.

**CONCLUSION**

For peptide and protein identification, the MetaProteomeAnalyzer software provides the possibility to search with different algorithms (X!Tandem, OMSSA, Crux, InsPect), each of which has its specific strengths. Combining the results guarantees a maximum reliability on the protein identification level. However, identification lists received as database search engine results are rarely sufficient for in-depth proteome analysis of microbial communities and merely serve as a starting point for further investigation. Indeed, a vast pool of knowledge about proteins is readily available in public online databases, but connecting protein search results with these resources typically remains a time-consuming and tedious manual task. MPA therefore automatically combines meta-information retrieved from a variety of sources to annotate protein identifications and thus categorizes the data in a meaningful, integrative fashion. An important benefit of the MPA is the flexible grouping of redundant protein identifications to meta-proteins. Instead of providing only one solution, the user can choose between different protein grouping methods or can even use a combination of multiple strategies depending on the required emphasis on protein diversity reduction or taxonomic resolution.

On the front-end side of the application, various visualization options facilitate addressing specific questions regarding proteins in relation to their functional and taxonomic information. The convenient usability of these features allows for unbiased data exploration. An additional degree of flexibility in categorizing and visualizing data is provided by the graph-based data handling and user-definable query system for dealing with complex questions.

The MPA pipeline has been developed with extensibility in mind regarding the addition of further sources of meta-information, for example, further ontology or pathway databases. Finally, while the MPA is geared toward processing metaproteomics data, the developed workflow also constitutes a highly useful tool for conventional proteomics data analysis.

**ASSOCIATED CONTENT**

Supporting Information

Application tutorial; data set description of microbial sample; SQL schema; meta-protein grouping rules; user-defined querying of data. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

Corresponding Authors

*E-mail: rapp@mpi-magdeburg.mpg.de. Fax: 49-3916110535.

*E-mail: lennart.martens@ugent.be. Fax: 32-92649484.

Author Contributions

¶T.M. and A.B. contributed equally to this work.

Notes

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

R.H. was supported by the German Environmental Foundation (DBU), Grant No. 20011/136. F.K. was supported by a grant of the Federal Ministry of Food, Agriculture, and Consumer Protection (BMELV) communicated by the Agency for Renewable Resources (FNR), Grant No. 22028811 (Biogas Biocoenosis). L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”), the IWT SBO grant “INSPECTOR” (120025), and the EU FP7 PRIME-XS project, Grant agreement No. 262067. The sample was kindly provided by the ABO Wind AG by Dipl.-Geograph Matthias Neuss.

**ABBREVIATIONS**

E.C., Enzyme Commission; emPAI, exponentially modified protein abundance index; LIIMS, laboratory information management system; MPA, MetaProteomeAnalyzer; MS, mass spectrometry; LC−MS/MS, liquid chromatography tandem mass spectrometry; SQL, structured query language; NSAF, normalized spectral abundance factor; FDR, false discovery rate; API, application programming interface

**REFERENCES**


Metaproteomics reveals major microbial players and their biogeochemical functions. *J. ISME* 2012, 6 (9), 1749–1762.


