The Peptonizer2000: graphical model based taxonomic identifications of metaproteomic samples

Tanja Holstein,[†,‡] Pieter Verschaffelt,[†,‡] Tim Van den Bossche,[†,‡] Lennart Martens,[†,‡] and Thilo Muth[*,¶]

†*VIB-UGent center for Medical Biotechnology, VIB, Belgium*

‡*Department of Biomolecular Medicine, Ghent University, Belgium*

¶*MF2, Robert-Koch-Institut, Berlin, Germany*

E-mail: mutht@rki.de

Abstract

Metaproteomics, the large-scale study of proteins from microbial communities, presents complex challenges in taxonomic inference due to sequence homologies between proteins within and across taxa. Commonly, taxonomic inference relies on heuristics, and few more advanced methods are available. We introduce the Peptonizer2000, a novel graphical model-based workflow designed to provide high-resolution taxonomic identifications of metaproteomic samples with associated confidence scores. This tool integrates peptide scores from any proteomic search engine with peptide-taxon mappings from the Unipept database, using advanced statistical modeling to enhance taxonomic resolution. We demonstrate the Peptonizer2000's accuracy and robustness through the analysis of various publicly available metaproteomic samples, showcasing its ability to deliver reliable probabilistic taxonomic identifications. Our results highlight the Peptonizer2000's potential to improve the specificity and confidence of

taxonomic assignments in metaproteomics, providing a valuable resource for the study of complex microbial communities.

Introduction

Microbial communities, also known as microbiomes, can be found in diverse environments such as ocean water,[1] biogas plants[2] and the human gut.[3] Recent technological and bioinformatic advances have enabled new discoveries such as the ability of microbiomes to use carbon monoxide in certain marine worms,[4] insights into the composition of certain environmental microbiomes,[5] the alteration of the gut microbiome depending on diet,[6] and the role microbiomes play in health and disease of humans[7] and animals.[8] Metaproteomics is a growing field that studies microbiomes by directly analyzing their proteins to investigate microbiome function and taxonomic composition.[9–11] Using technologies from traditional proteomics, tandem mass spectrometry (MS/MS) is employed to analyze proteins in biological samples. However, microbiome samples are more complex to understand and handle, posing challenges in experimental and bioinformatic setups.[12,13]

One challenge is figuring out which microorganisms are in the sample. This is complicated because proteins often share similarities in their sequences and can be linked to multiple groups of organisms.[14] In very diverse samples, it can be tough to identify the specific species or even families of microorganisms present.

Several bioinformatic workflows that identify taxa and functions in microbiome samples have been developed. Among the most popular are iMetaLab,[15] the galaxy framework[16] and Unipept.[17] To identify the taxa present in a microbiome sample of unknown origin, these workflows rely on heuristics such as peptide-taxon match counting to identify the taxa present. Very commonly, taxa are identified as present if at least two or three unique peptides (mapping to the taxon in question) are detected.[18–20] When a peptide cannot be assigned to a taxon at a species level, it will then be mapped back to the lowest taxonomic level

it is specific to, reducing the taxonomic resolution of the analysis. This taxonomic level is referred to as the lowest common ancestor or LCA.[21] The LCA approach is key to the taxonomic identification done by Unipept. Over the past years, the number of taxa and associated proteins in the publicly available databases like Uniprot has steadily increased.[22] This has been shown to overall decrease the specificity of peptides at lower taxonomic levels thus reducing the taxonomic resolution of the LCA approach.[17]

Addressing these issues requires more sophisticated taxonomic identification algorithms. MiCiD[23] is such an integrated metaproteome analysis tool with robust e-value calculation for taxonomic identifications. However, it includes its own database search for identifying mass spectra, intrinsically bound to the taxonomic assignment. It is thus not compatible with the outputs of other very popular search engines such as X!Tandem[24] or rescoring pipelines such as MS2Rescore[25] and is limited to data-dependant acquisition collected mass spectra.

Unipept, on the other hand, simply takes lists of peptides as input. These can be generated form any search engine results. It is however unable to take into account scores calculated for peptides by the search engines and is limited to counting the number of peptides per LCA to identify present taxa.

None of the aforementioned software solutions offer both advanced statistics, straightforward usability, advanced visualization options and the possibility for the user to submit results from any preferred search engine, thus potentially applicable to DDA and DIA data. The potential inclusion of DIA is of interest as it has become more popular in recent studies.[18,26]

We here introduce the Peptonizer2000: A graphical model-based workflow for taxonomic identification of metaproteome sample with associated confidence scores. We previously developed PepGM,[27] a graphical model based workflow for the taxonomic identification of single virus proteome samples. Strong protein sequence homology between different viral strains can make accurate, strain-level identification difficult, an issue analogue to the one faced when identifying the species present in microbiomes. We have shown that the use of

a more advanced statistical model allows for a more accurate taxonomic identification. The Peptonizer2000 integrates statistical modelling for taxonomic identification, taking into account peptide scores computed by proteomic database search engines, with the peptide-taxon mapping information provided by Unipept, to offer high resolution taxonomic identification of microbiome samples with probability scores associated to each taxon.

We show that the Peptonizer is able to compute taxonomic confidence scores on user-selected taxonomic levels for various metaproteomics samples using different background databases. We compare the Peptonizers probabilistic identifications to Unipepts number-of-peptides-based identifications, describing how the taxas scores provide valuable additional evidence to discriminate between truly present taxa and random taxon-peptide matches.

The Peptonizer2000 workflow

In the following section, we describe the Peptonizer2000 workflow in detail. We start with an overview of the workflow followed by a more detailed description of the new or adapted workflow steps.

The Peptonizer has several options for the files that can be provided as input. The first is a set of MS/MS spectrum files (.mgf format) and a reference database (.fasta) to use for the proteomic search. If the user wishes to use their own database search engine, a set of peptide-spectrum matches in percolator output format (.pout) can be provided. A third option is to provide a .csv file containing peptide-spectrum matches and corresponding scores, where the scores have to represent a probability for a peptide to truly be present in the sample. This set of input formats should allow, for example, for users to apply the Peptonizer2000 workflow to search results from both DDA and DIA acquired peptides. An overview of the Peptonizer2000 workflow in shown in Figure 1. It is composed of the following steps: (1) standard proteomic database search, (2) query of all candidate peptides in the

Unipept database and collection of candidate taxa based on weighted PSMs (3) graphical model construction, belief propagation and grid search through graphical model parameters (4) parameter evaluation and (5) results output and visualization.

The steps were adapted from our previously developed taxonomic identification tool PepGM as mentioned previously. Should a .pout or .csv file be provided as input, step(1) is skipped. Snakemake[28] is used as a workflow management system while all scripts are written in python and developed and tested for LinuxOS and python 3.10. All additional python packages used are listed in the supplementary materials.

The Peptonizer2000 source code is open source and is available under `https://github.com/BAMeScience/Peptonizer2000`.

Now, we describe the Peptonizer2000 steps in detail. In particular, we highlight any changes or adaptations that were made to the workflow compared to the already available PepGM.[27]

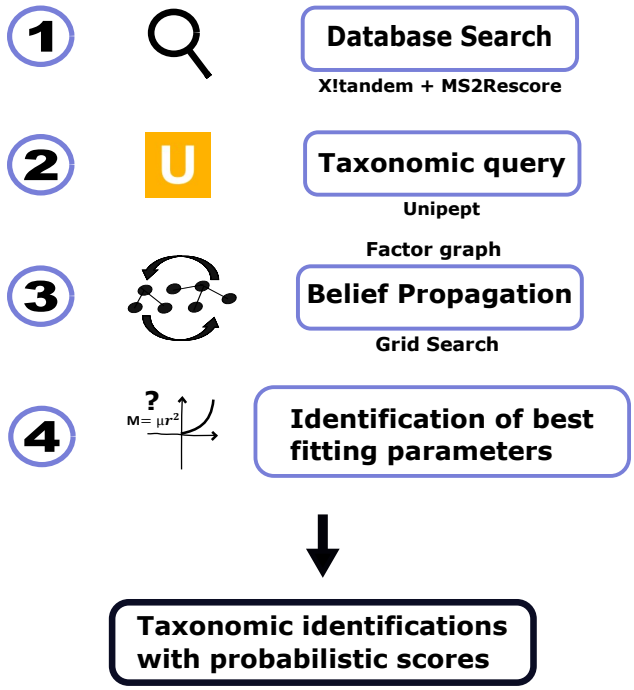

Figure 1: Peptonizer Workflow

(1) Proteomic database search

MS/MS spectra are searched against the reference database provided by the user using a combination of X!Tandem[24] and MS2Rescore.[25] This yields a list of identified peptide-spectrum matches (pout format) with corresponding scores. The e-value computed by MS2Rescore will be used for the following probabilistic computations.

As previously mentioned, this database search step is optional. Alternatively, the user can choose to provide scored peptide-spectrum matches as mentioned above, and can therefore use any preferred database search engine.

(2) Querying Unipept to retrieve the taxonomic annotations of peptides

All peptides above a user-defined false-discovery rate (FDR) are aggregated with their corresponding number of PSMs and scores. These peptides are then queried in the Unipept database.

To this end, the Unipept database was equipped with a new API endpoint. This endpoint returns all taxa a peptide is associated to through a Uniprot[22] protein. This endpoint does not compute or return a LCA taxon, as was previously the case. Three queries are always executed in parallel and peptides are queried in chunks of 10 to avoid server overload.

Additionally, to speed up the peptide queries, peptides mapping to more than 10000 proteins are excluded from the search. Query times increase with the number of proteins to look up in the Unipept database, and peptides mapping to that many proteins are likely part of so-called housekeeping proteins,[29] adding little taxonomic information. A brief investigation into the LCA of the hereby excluded proteins is shown in the supplementary materials in section 'Taxonomic rank of peptides and proteins excluded by the Peptonizer' and confirms this.

If there is prior knowledge about potentially present microorganisms in a sample, the user can choose to restrict the Unipept query to taxa of interest only. The taxa of interest can be specified at any taxonomic level, by specifying their corresponding taxid (NCBI unique taxon identifiers). For example, if the taxid to be queried is 2, the Unipept query will include only bacterial references proteomes. In all of our subsequent analyses, we never restrict the taxonomy of the Unipept query - corresponding to situation where no prior knowledge would be available to researchers.

Querying Unipept results in a Unipept response .json containing the taxonomic information linked to the peptides, their scores and PSM number.

Selection of candidate taxa

The peptides from a metaproteomic sample will map to a very wide range of taxa. This would result in a graph with well over 1000 potentially present organisms and very long computation times. We therefore restrict the number of taxa included in the graph. As a standard, we set this number to 150, but the number of taxa to be included is user-defined through the configuration file. Every taxon with at least one unique peptide will be included in the graph, even if the number of included taxa then exceeds 150.

The number of peptides detected for a taxon that is present has a very wide range. It depends both on organism size and number of organisms present or their share of the protein mass in a sample. For instance, in one MS run for the very commonly used, lab-assembled SIHUMIx mixture,[30] 5 PSMs could be detected mapping to *Lactobacillus plantarum*, while over a 1000 map to the much more abundant *Bacteroides thetaiotaomicron*. Both taxa are present in SIHUMIx. We aim to address this issue by assigning a high weight to unique PSMs. This makes sure that the number of spectra matching to a peptide is taken into account as well as the taxonomic information (linked to a peptide's sequence degeneracy) that peptide carries.

The resulting formula to compute the weight W_T of a taxon is the following:

$$W_T = \sum_i \frac{\#PSM_i}{Deg_i} \tag{1}$$

Where i stands for all peptides mapped to taxon T, $\#PSMS_i$ is the number of PSMs per peptide i and Deg_i is the degeneracy of the peptide i, i.e. how many taxa it was mapped to. All candidate taxa and their corresponding, scored peptides are exported as a .csv file used as basis for the graphical model.

(3) Assembly of the graphical model, belief propagation and grid search

Assembly of the graphical model

The construction of the graphical model, which is a factor graph, uses the same algorithm as PepGM. All peptide and taxa from step (2) are included in the graph. Two categories of nodes represent peptides and taxa respectively. An edge is drawn between a peptide node and a taxon node if a peptide is part of a taxons proteome. Exactly as described in,[27] factor nodes based on the noisy-OR model and convolution tree nodes are added to the graph. Together, the graphical model represents the probability distribution of all peptides and taxa. Note that this probability distribution represents an approximation, as not all taxa are included in the graph, and includes the same assumptions regarding peptide emission and detection probabilities as PepGM. The noisy-OR model introduces 3 parameters. Very briefly, these are:

- α : probability for a peptide to be observed given the presence of its parent taxon
- β : probability for a peptide to be randomly (falsely) observed or have an erroneous connection to a taxon (which is not actually present in the sample)
- γ : prior probability for a taxon to be present

We have discussed these parameters more extensively for PepGM.[27]

With the Peptonizer2000, we provide the additional possibility to regularize the probabilities for a peptide to have n parent taxa present in the noisy-OR model. This regularization has been applied for protein inference previously.[31] This models a probability distribution for samples where a very broad taxonomic range was queried compared to the expected number of present taxa. The user can choose to turn the regularization on or off. As for metaproteomics samples, generally, the reference databases contains many more taxa than are actually detected or present in the sample, we generally recommend to keep the regularization active.

Inference algorithm and grid search

To compute the marginal probabilities of the taxa, we use the belief propagation algorithm.[32] To address the large size of the graph, we implemented a version of belief propagation termed 'zero-lookahead' by it's creators.[33] This version of belief propagation was shown to have nearly linear order compared to graph size and represents only a slight approximation of the global probability distribution. Additionally, we perform a grid search through the 3 model parameters introduced by using the noisy-OR model. To reduce search space, we set the prior probability $\gamma \in [0.1, 0.3, 0.5]$. The peptide emission probability – meaning the probability for a peptide to be detected should a parent taxon be present, is set to $\alpha \in [0.85, 0.9, 0.99]$, while the error probability,$\beta \in [0.5, 0.6, 0.7]$ by default. This can be altered by the user. We will discuss the basis of this parameter selection and their interpretation in the results section. The exact parameters best suited for a specific sample are selected in the next step.

(4)Selection of most sample appropriate parameter range

For each combination of model parameters α, β and γ, a set of marginal probabilities is computed for all taxa included in the graphical model. Now, the optimal set of parameters for the sample at hand needs to be selected.

The evaluation of the best fitting parameters relies on a comparison between the results of the belief propagation and a list of weighted taxa that was used as input for the graphical model. We denote the scored list L_S and the weighted list L_W in the following. Both lists of taxa are sorted by score and weight, respectively.

Since the weights attributed to each taxon rely on the number and uniqueness of PSMs per taxon, the weighted taxon list L_W gives a first indication about the taxa that should be identified as present. It has one central flaw: neighbouring taxa, such as three or more very similar bacterial species, may all be present high up in L_W. However, we aim to select a parameter range that would be able to resolve between these taxa. The solution we propose is to cluster the list L_W by taxon similarity. The clustering procedure is analogous to taxonomic clustering previously described in,[23] with the exception that the similarity between two taxa is based on their detected peptides only and not their whole theoretical peptidome. This choice ensure that the clustering is closer to the currently analyzed sample, where taxon abundances and number of detected peptides may vary greatly. The taxon with the highest number of detected peptides among similar taxa will always be a cluster head. The threshold similarity for taxa to be in the same cluster is set to 0.9, this choice is arbitrary. However, since we want to cluster similar taxa, it is clear that this measure should be well above 0.5, but not equal to 1. Slight variation of up to 0.1 lower similarity do not results in differences of more than one or 2 taxa in the clustered taxon list.

After the taxon clustering, we obtain a simplified version of the weighted taxon list L_W, that only contains the weight-sorted cluster heads.

Both lists are then ready to be compared. For the comparison, we need to take into account the following: taxa listed in the results of the Peptonizer L_S might be absent from the clustered L_w, we therefore need a comparison that allows for non-conjointness. Additionally, the higher weighted and scored taxa are more important for our list comparison than the lower ranked and scored ones: higher ranks in the list should be higher weighted than low.

As measure for list comparison, we therefore choose Rank-biased-overlap[34] (denoted RBO), which fulfills the above mentioned criteria.

The parameter range selected for a specific sample will be the one maximizing the rank-biased-overlap between the clustered L_W and the scored results L_S and minimizing the entropy of the marginal probability distribution. Entropy, denoted S, was already used as a criterion for PepGM[27] and its consideration aims to maximizes the information content of the obtained distribution. The final metric, that the chosen parameter range for a sample should maximize, is:

$$M = RBO \cdot S^{-2} \tag{2}$$

Results output and visualization

The results of the Peptonizer2000 are output in various formats. A .csv file containing all taxa with their corresponding score is the most basic output level. A bar chart providing a rapid overview of the results is generated as a .png, together with a bar chart of how each parameter set performed according to the rank-biased-overlap and entropy metric.
A visually appealing treeview of the results can also be generated in both .html and .svg format.

Materials and methods

We use several publicly available metaproteomic samples to evaluate the Peptonizer2000. All samples are available through the PRIDE database.[35]
The first set of samples was obtained during the CAMPI study.[36]
The first sample type we choose to analyze is the SIHUMIx sample.[30] SIHUMIx was developed as a model community for the human gut microbiota and covers the three dominant genera in human feces, *Firmicutes*, *Proteobacteria* and *Bacteroidetes*. It contains eight

microbes: *Bacteroides thetaiotaomicron*, *Anaerostipes caccae*, *Escherischia coli*, *Lactoplantibacillus plantarum*, *Clostridium butyricum*, *Thomasclavelia ramosa*, *Blautia producta* and *Bifidobacterium longum*. It is very commonly used in metaproteomics benchmarking. Analogous to the original CAMPI publication, we use the samples s03, s05, s07,s08 and s11 for our investigation and will refer to them using the identifiers s03-s11. Since each sample was acquired with different laboratory workflows, they offer slightly different coverage of the proteomes contained in the samples.

The second sample type taken from the CAMPI study is a fecal sample - here, the taxonomic 'ground truth' is unknown. We choose the fecal sample denoted as F07 for our analysis.

The second set of samples we analyze is a more complex, lab-assembled mixture of up to 32 different microorganisms.[37] They are available on PRIDE under the identifier PXD006118. We select samples from three types of communities for our analysis: these were assembled with equal cell amount per taxon (denoted 'C1'), equal protein amount per taxon (denoted 'P1') and uneven composition (denoted 'U1').

We tested 4 different reference databases: one database is tailored to the SIHUMIx sample and was available through CAMPIs corresponding pride repository. The second database is the integrated gene catalog (IGC) gut microbiota reference database,[38] which contains representative sequences aiming to cover most gut microbes. The third database is tailored to the 32 potentially present microorganisms in the complex mock community and was provided alongside the experimental data.[37] These three databases are concatenated with a database of contaminants,[39] so that peptide hits matching the contaminant database will be filtered out before taxonomic analysis.

Finally, as a fourth reference, we use the UniREF50 database.[40] This is a clustered set of sequences from all of Uniprot, aiming to conserve coverage of the sequence space but reducing redundancy.

For the proteomic database search, we fixed the search parameters in accordance with the original sample publications. These were the same settings for all samples analyzed: a spe-

cific trypsin digest with a maximum of two missed cleavages; mass tolerances of 10.0 ppm for MS1 and 0.02 Da for MS2; fixed modification: Carbamidomethylation of C (+57.021464 Da); variable modification: Oxidation of M (+15.994915 Da); Peptide length between 6 and 50 amino acids, and charge state +2, +3, or +4. Additionally, the false discovery rate (FDR) was set to 1%.

To perform the Unipept analysis, we use the Unipept command line tool '–pept2lca' with as input a .txt file of all peptides that were used as input for the Peptonizer. Since Unipept outputs the LCA for the queried peptides, we filter the Unipept response on peptides unique to taxa at species level and aggregate the number of unique peptides per species to obtain Unipept taxonomic identifications at species level based on unique PSMs. For the analysis of higher taxonomic levels, we use the same approach, mapping unique peptides at lower taxonomic levels back to ancestors if necessary.

All code used for downstream analysis is also available at `https://github.com/BAMeScience/Peptonizer2000`. The code used to generate the treeview graphics of the Peptonizer results is available at `https://github.com/pverscha/peptonizer-visualizations`.

Results

To showcase the added benefit of calculating confidence scores for identifying taxa in metaproteomic samples and to showcase the Peptonizers accuracy in doing so, we analyze a selection of publicly available samples with Unipept alone and with the Peptonizer.

Taxonomic identification results for samples with known composition and varying complexity

As a first test, we use only lab assembled mixtures. This ensures that the actual taxonomic composition of the sample - a ground truth for the species present - is known.

Taxonomic composition results for the SIHUMIx sample

As described in the materials section, first, we choose the samples s03, s05, s07 and s11 of SIHUMIx from the CAMPI study. As a reference database, we use a sample-specific SIHUMIx database, provided alongside the MS/MS data in pride, analogously to the original CAMPI study. In this database, the included species of the genus *Blautia* is *Blautia pseudococcoides*, as this was the reference proteome available at the time. In the following, we therefore also accept *Blautia pseudococcoides* as taxonomic identification result.

Figure 2 shows the analysis results for the top scoring taxa for the sample s07. The Peptonizer2000 identifies all present taxa as top scoring taxa correctly, with probability scores between $p = 1$ and $p = 0.9$. All present taxa also correctly appear among those with the highest number of unique peptides. The results also demonstrate several notable attributes of the Peptonizer and Unipept.

The Unipept results show the very high range in the number of detected peptides (between

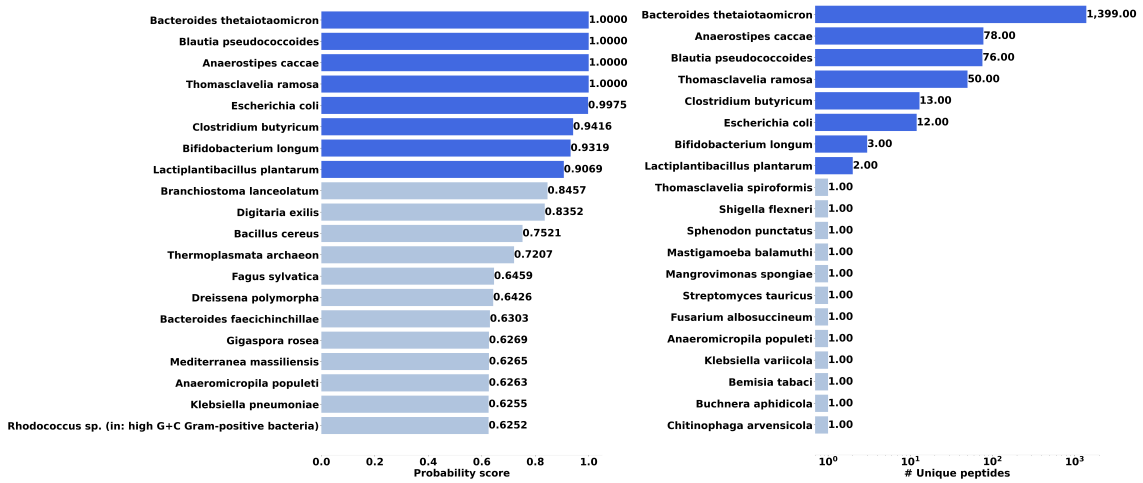

Figure 2: Results for the taxonomic identification of the SIHUMIx samples s07 using the Peptonizer2000, left, with corresponding probability scores, and using unique peptides from Unipept, right. The results show good overlap.

1399 and 2) for the taxa present in the sample. From the Unipept results alone, selecting the present taxa with arbitrary criteria like 'a minimum of 2 unique peptides per taxon' is hard to justify, even though in the present case, it would lead to correct identifications. This

type of cut-off is very common in metaprotoemic workflows.[18–20,41,42]

The Peptonizer results attribute all highest scores to the correct taxa. Since they represent a probability distribution, introducing a 'cut-off' probability for present taxa does not make sense mathematically, but will often be necessary for practical reasons. Here, looking at the results from Unipept and the Peptonizer conjointly, a posterior taxon probability of about $p = 0.9$ could be suggested. Additionally, looking at the peptonizer and UNipept results conjointly will lead to the best results: The 8 taxa with the highest number of unique peptides also obtained high Peptonizer scores, one can therefore safely identify them as present. None of the taxa with one unique peptide appear in the higher scored taxa of the Peptonizer results, they can therefore be identified as absent with a very high probability. The next higher scoring taxon in the Peptonizer results is *Branchiostoma lanceolatum*. In the Peptonizer graph, this species has 51 shared peptides attributed to it, which is why it still obtains a relatively high probability score of ≈ 0.85. However, looking at the Unipept and Peptonizer scores together, it can safely be identified as absent.

Figure 3 shows another example of Peptonizer and Unipept results, this time for the SIHUMIx sample s11. Notably, the Peptonizer wrongly identifies the taxa *Candidatus Woesearcheota archaeon* and *Candidatus Woesearcheota archaeon* as present. They are taxa that have one unique peptide and many shared peptides with actually present taxa. Most importantly, the *Candidatus* in the taxon name refers to putative, yet uncultured eukaryotic taxa, and the records associated to *Candidatus* taxa are known to be error prone and chaotic.[43] Researchers can thus easily identify it as an erroneous classification. Regarding the rest of the high scoring taxa from the Peptonizer together with the unique peptides identified by Unipept, the same remarks as for the sample s07 remain valid: for the present sample, the Peptonizer probability scores justify a cutoff of three unique peptides per taxon to identify it as present. Additionally, again, a posterior probability o $p = 0.9$ seems to reliably identify present taxa. The other taxa in the Unipept results, even those with two unique peptides

per taxon, obtain lower probability scores.

Similar conclusions can be drawn from the Peptonizer and Unipept results for the samples s03, s05 and s08. These are shown in the supplementary materials in figures 1 to 3.

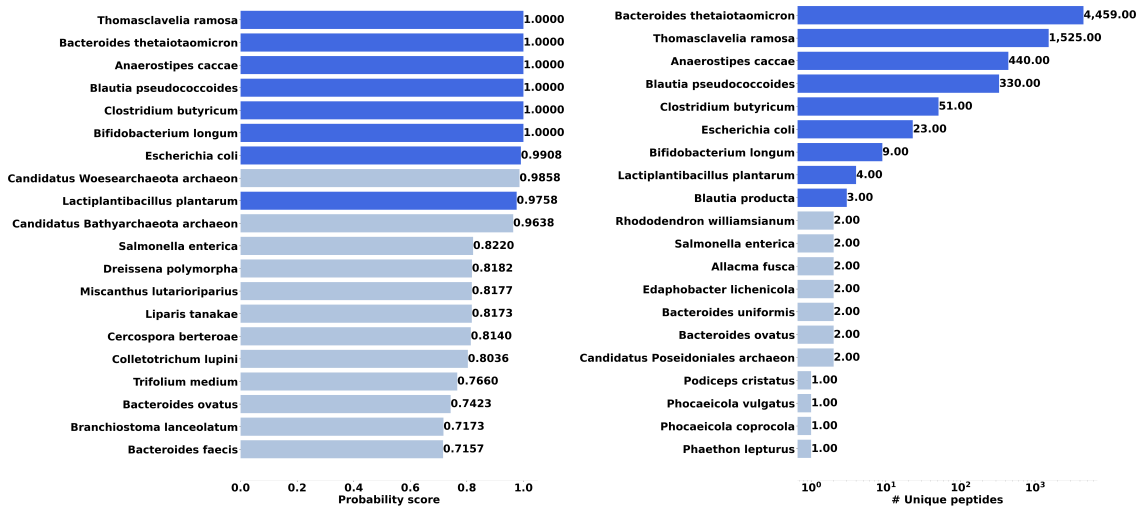

Figure 3: Results for the taxonomic identification of the SIHUMIx samples s11 using the Peptonizer2000, left, with corresponding probability scores, and using unique peptides from Unipept, right. Except for *Sus scrofa*, the results show good overlap.

Taxonomic composition results for more complex samples

To demonstrate how the Peptonizer performs for more complex samples, we use a lab-assembled mixture of up to 32microorganisms from a previous metaproteomic study.[37] We analyze the samples U1, C1 and P1, who have been respectively constructed with uneven taxonomic abundance, protein and cell number (U1), equal cell amount (C1) and equal protein amount(P1).

Figure 4 shows the Peptonizer identification results for the sample U1 in a treeview diagram. All taxa that obtained a probability score above 0.8 are displayed. Additionally, Figure 5 shows the Peptonizer attributed scores for all taxa that scored above 0.8 together with the 30 taxa that were attributed the highest number of unique peptides using Unipept.

Out of the 32 potentially present taxa, 28 were included in the sample. Of these, 5 are phages. These are not detected, likely due to their very low protein amount and filtering steps currently applied by Unipept. Additionally, viral and therefore phage taxonomic annotations of the reference databases tend to be much less complete and less well curated than those for bacteria or eukaryotes.[44] Out of the other 23 microorganisms, 21 are correctly identified. The two missing taxa are *Roseobacter Sp. 199* and *Agrobacterium tumefaciens*. Some erroneous taxa also received scores above 0.8. For most, they are closely related to present species: *Rhizobium johnstonii* was wrongly identified as present along with the correct *Rhizobium leguminosarum* from the same genus. The same observation can be made for the wrong identification of the two *Pseudomonas* species *indicatrix* and *lactitubi*. The Peptonizer is not able to resolve between the species inside these geni based on the detected peptides.

The comparison between the Peptonizer and Unipept results in Figure 5 again reveals the added benefit of their results combination: the Peptonizer scores can help justify which taxa

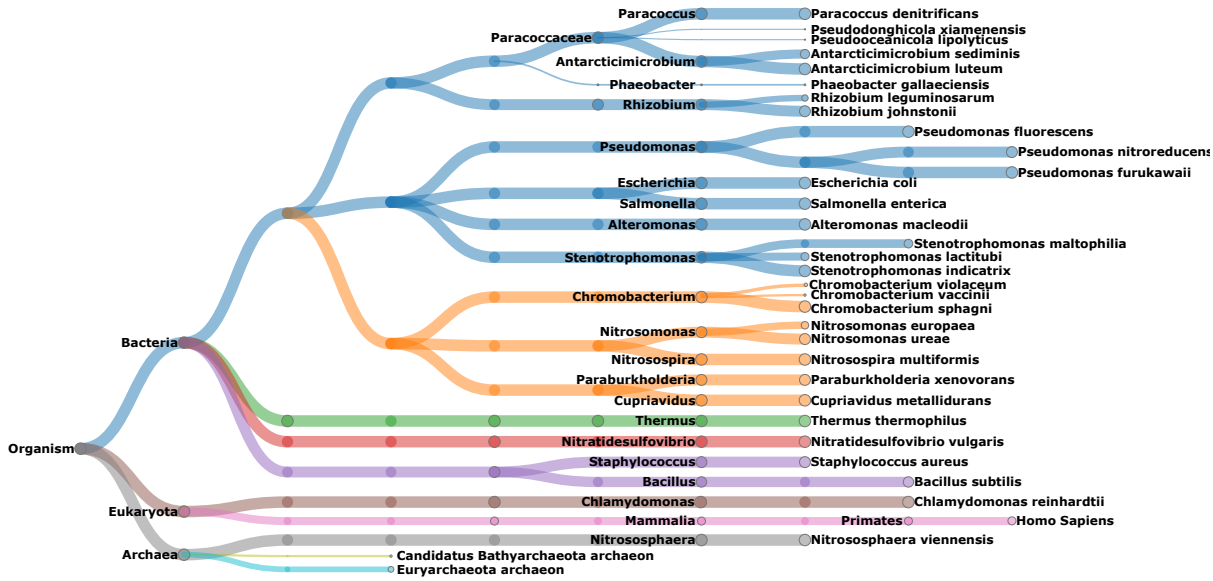

Figure 4: Peptonizer Results shown as taxonomic tree view for a lab assembled sample of 28 different species. Node sizes are proportional to probabilities attributed by the Peptonizer.

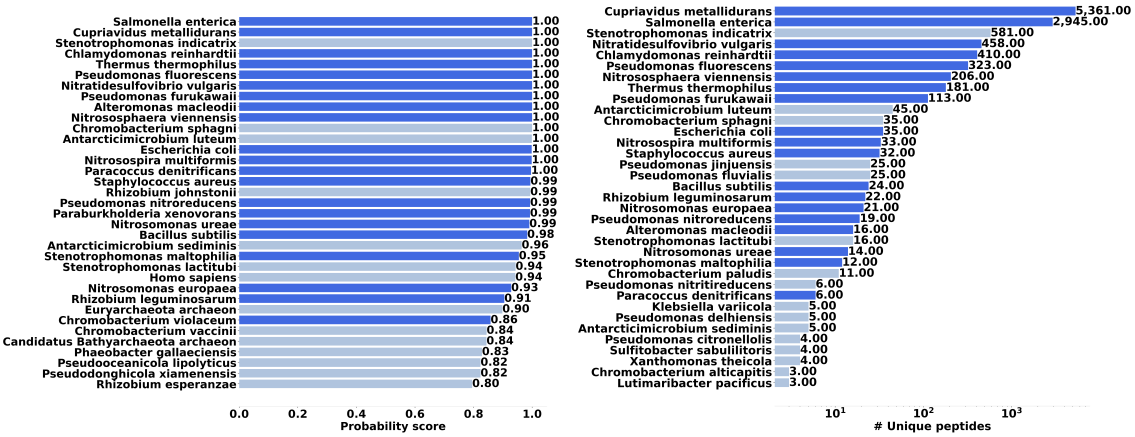

Figure 5: Results for the taxonomic identification of the lab assembled community U1 using the Peptonizer2000, left, with corresponding probability scores, and using unique peptides from Unipept, right.

with unique peptides are accepted as present or absent, especially regarding the vast range of unique peptides identified between present taxa (between 5000 and 6). In the present sample, the Peptonizer scores can help identify *Paracoccus denitrificans* as present but *Pseudomonas nitrireducens* as absent even though both taxa have 6 unique peptides attributed to them. Some false positive identifications remain - for example, *Antarctimicrobium sediminis* obtains a high probability score of 0.96 and 5 unique peptides, even though it is not present in the sample. Two archeota species are also false positives in the Peptonizer results. One of them contains *Candidatus* in its name, which, as previously mentioned, might hint towards errors in the proteome references. Since archeae seem to be frequently misidentified, it could be helpful to include an optional filtering for 'archeae' taxa in future updates.

The taxonomic analysis of samples C1 and P1 are shown in the supplementary materials. While the number of correct identifications is slightly lower, 17 for community C1 and 14 for community P1, the observations regarding the synergy of unique PSMs computed by Unipept and Peptonizer scores remain valid.

Search results using untailored reference databases

To investigate the performance of the Peptonizer when untailored reference databases are used, we search the SIHUMIx sample s07 against two other proteome reference databases.

First, we search against the IGC reference catalogue, using the fasta file provided in the CAMPI pride repository. As this file contains the reference proteomes of many genera commonly found in the human gut, it contains the genera present in the SIHUMIx sample, but may not contain a reference for the exact species present in SIHUMIx. In analogy with the presentation of the previous results, we also show the taxonomic identifications based on unique peptides at species level obtained through Unipept alone. Three of the species present in lower amount in SIHUMIx are detected neither by Unipept nor the Peptonizer, these are *Lactoplantibacillus plantarum*, *Clostridium butyricum* and *Bifidobacterium longum*. Because of the use of a larger, less well fitting database, no peptides unique to these species were detected anymore. Instead, for each, so few peptide hits are detected that they are not included in the Peptonizer graphical model. For the five taxa that are detected, the Peptonizer probability scores, together with the number of unique peptides detected, can allow researches to make correct conclusions about the sample's composition: All five correct taxa obtained the highest scores using the Peptonizer. The next highest scoring taxon, *Blautia obeum*, does not appear in the top taxa of the Unipept analysis and can therefore be excluded. Using Unipept alone, it would have been less clear which taxa are present.

To simulate a situation where there is no prior knowledge about a sample at all, we search the samples s07 from CAMPI and the sample P1 from the more complex lab-assembled mixture against UNIREF50. Importantly, with a reference database this large and untailored to the sample at hand, the identification rate of peptides can be expected to drop drastically, which in turn will lead to worse taxonomic identification results. In this analysis, we included an analysis with a peptide FDR of 5% additionally to the common 1%, to investigate whether

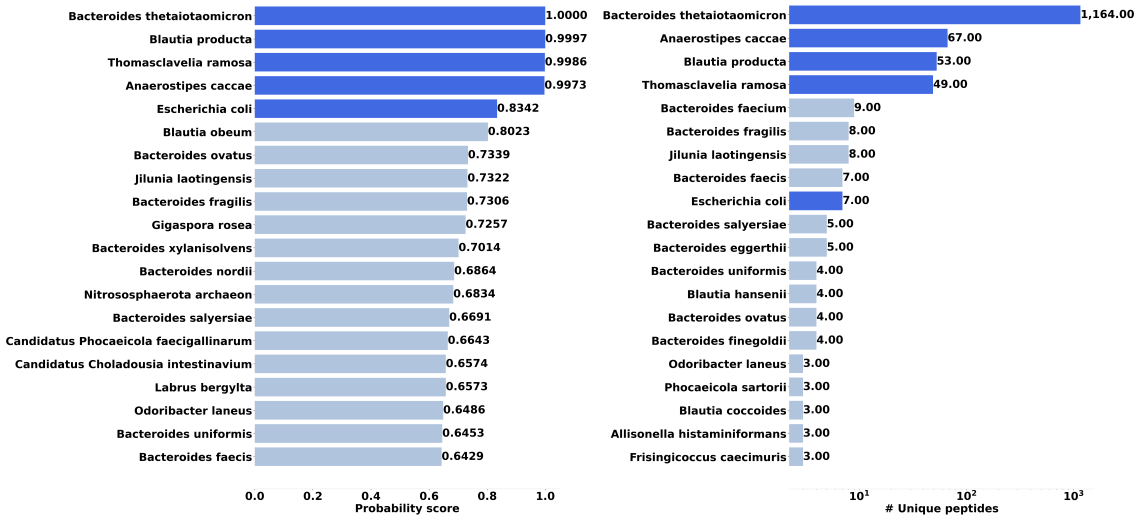

Figure 6: Results for the taxonomic identification of the SIHUMIx sample s07 searched against the IGC, using the Peptonizer2000, left, with corresponding probability scores, and using unique peptides from Unipept, right.

this might improve the taxonomic identification results. The results for these analysis are shown in the supplementary materials in figures 8 to 11. Overall, as expected, the use of the very large reference database is detrimental to the taxonomic identification of the samples. For SIHUMIx, a maximum of 4 correct taxa (*Bacteroides thetaiotaomicron*, *Anaerostipes Caccae*, *Thomasclavelia Ramosa*, *Blautia producta*) appear in the 20 highest scoring results. Additionally, many false positives appear. While the identifications are slightly better for the Peptonizer when a FDR of 5% is used, it is worse for the Unipept analysis. This is be due to the Peptonizers ability to take into account the peptide scores - peptides with lower confidence will contribute to the taxonomic identification, but the information about their lower confidence will contribute towards the computed probability scores. However, overall, no clear taxonomic attributions can be made using UNIREF50.

For the sample P1, a maximum of 9 correct taxa appear among the 35 highest scored taxa. The major of taxa obtaining a probability score above 0.9 are truly present in the sample, however these are only 6 out of 23 taxa, again highlighting the issue that database size poses in metaproteomics in general.[45]

Analysis of a fecal sample

We now discuss the Peptonizer results for a fecal sample. As mentioned above, we use the fecal sample F07 from the CAMPI study. As for the CAMPI study, this fecal sample is searched against the IGC gut reference database. We analyzed the sample at three taxonomic resolution levels: species, genus and family. Figure 7 shows the Peptonizer identification results for the families present in the sample F07 that received a probability score higher than 0.6. 22 bacterial families are detected. Overall, the results agree with previous analysis of fecal samples in CAMPI[36] and other studies on the human gut microbiome.[46,47] This demonstrates the Peptonizers abilities to compute taxonomic probability scores at higher taxonomic levels. The bar plot of the Peptonizer scores together with their comparison to the Unipept results

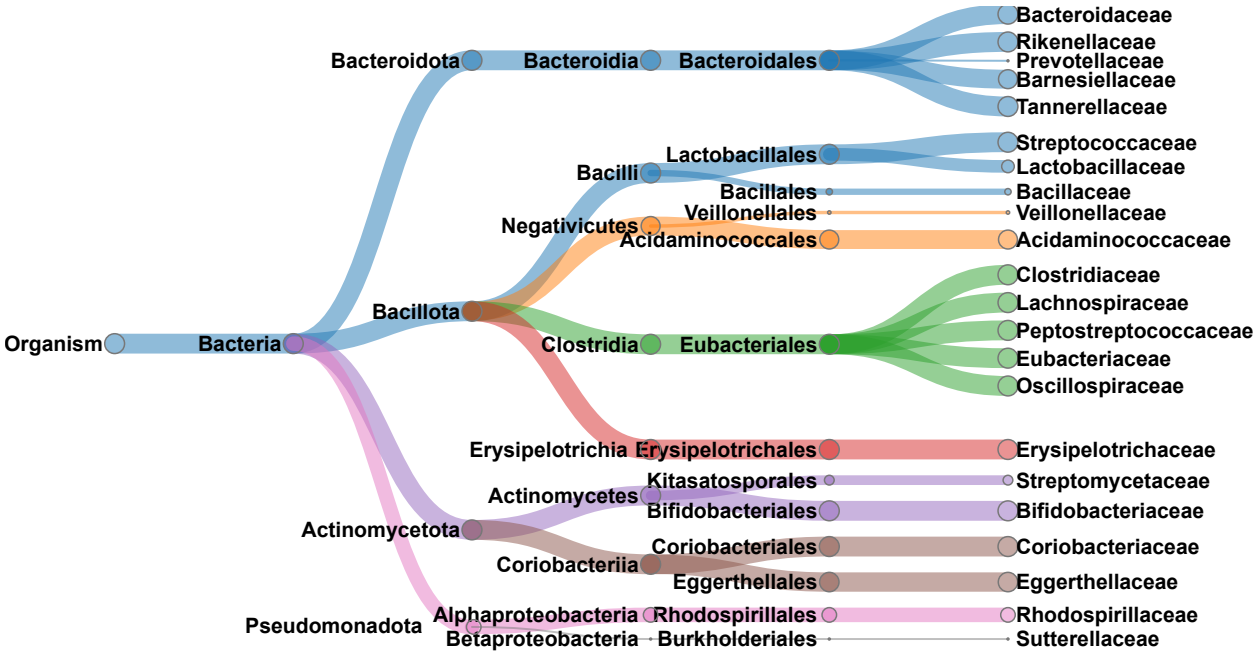

Figure 7: Peptonizer Results shown as taxonomic tree view for a fecal sample taken from the CAMPI study. Node sizes are proportional to probabilities attributed by the Peptonizer. The results are in good agreement with previous analyses of fecal samples.

is shown in the supplementary materials in Figure 6. The range in the number of unique peptides per family is very large - from 1 to 8190. All families with a very large number of unique peptides are also attributed very high probability scores by the Peptonizer. For families with lower numbers of taxonomic attributions, the Peptonizer scores can hint towards

the certainty of their presence. For example, the family *Suterellaceae* has 38 unique peptides detected, and a probability of presence of 0.63, and is therefore more likely to be present, which overlaps with previous analyses of this sample.[36] *Flavobacteriaceae* obtain a score of 0.58, even though only one unique peptide is detected. As this family of bacteria has been isolated from fecal samples before,[48] they might truly be present in the fecal sample F07, but no further conclusions can be drawn, since no ground truth is available. Figures 7 and 8 in the supplementary materials show the Peptonizer and Unipept results at genus and species level respectively. Many species are identified as present by the Peptonizer, however, human gut microbiomes are infrequently analyzed with comprehensive species level resolution. We therefore focus on commenting on the genus level results.

At the genus level, most geni attributed a high number of unique peptides by Unipept also obtain high Peptonizer scores. To allow for better comparability, Figure 7 in the supplementary materials shows the Unipept results order in descending order of probabilities attributed by the Petonizer. The 60 highest scoring geni are shown. Notably, certain geni that are known to be present in the human fecal microbiome, but only have few or no unique peptides, still obtain high Peptonizer scores. This is the case, e.g., for *Anaerobutyricum* (no unique peptides , Peptonizer score of 1.0),[49] *Fusicatenibacter* (1 unique peptide, Peptonizer score of 1),[50] *Flavobacterium* (1 unique peptide, Peptonizer score of 0.83),[51] *Simiaoa* (no unique peptides, Peptonizer score of 0.95) and *Waltera* (no uique peptides, Peptonizer score of 0.59).[52] On the other hand, for few geni that have a high number of unique peptides detected, the Peptonizer scores are slightly pessimistic. This is the case, e.g., for *Butyricoccus*,[53] which is attributed 84 unique peptides but obtains a Peptonizer score of 'only' 0.43. This showcases the advantage of jointly interpreting Unipept and Peptonizer results for avoiding false negative and positive taxonomic identifications. As a side note, many likely erroneous taxonomic identifications of the Peptonizer belong to the *Candidatus* category, which, as previously discussed, likely contains many wrong references.

Parameter choices for the graphical model

For all analyses, the Peptonizer was run for the same ranges of parameters α, γ and β. The meaning of this parameters is described in detail previously. For every search, $\alpha \in [0.7, 0.8, 0.9, 0.99]$, $\beta \in [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ and $\gamma \in [0.1, 0.3, 0.5]$. A grid search then identifies the best fitting parameter set for each sample. The parameter set selected for each sample is shown in table 1. The parameter α represents the probability for

Table 1: Parameters of the noisy OR conditional probability tables, α, β and γ, selected as best fitting for each sample.

Sample	Origin/Composition	Reference database	α	β	γ
S03	CAMPI/SIHUMIx	tailored	0.99	0.7	0.5
s05	CAMPI/SIHUMIx	tailored	0.99	0.7	0.5
s07	CAMPI/SIHUMIx	tailored	0.99	0.7	0.5
s07	CAMPI/SIHUMIx	IGC gut	0.99	0.9	0.1
s07	CAMPI/SIHUMIx	UNIRef50	0.99	0.7	0.3
s08	CAMPI/SIHUMIx	tailored	0.99	0.6	0.5
s11	CAMPI/SIHUMIx	tailored	0.99	0.6	0.5
s11	CAMPI/SIHUMIx	tailored	0.99	0.6	0.5
F07	CAMPI/fecal	IGC gut	0.99	0.6	0.5
U1	Kleiner/28Mix	tailored	0.99	0.7	0.5
P1	Kleiner/28Mix	tailored	0.9	0.5	0.1
C1	Kleiner/28Mix	tailored	0.99	0.7	0.1

a peptide to be detected if the parent taxon is present. Since many peptides are detected in metaproteomic samples, it makes sense that $\alpha \geq 0.9$ was selected for all samples: if a parent taxon is present, it is very likely that a peptide will be observed. This choice of parameter is also in line with parameter ranges identified as optimal in other studies using a similar graphical model: for viruses, where the probability to detect peptides is relatively low, the optimal α was found the be small[27] while for more general protein inference, α was identified to be much higher.[31]

The parameter β represents the probability for a peptide to be detected at random, but can also be interpreted as the probability for a connection between a parent taxon and a detected peptide to be erroneously present. $\beta \geq 0.5$ was identified for all samples, meaning the model

assumes there is a high probability for a wrong peptide or a wrong connection between the peptide and its parent taxon. This observation is in good agreement with the setup of our metaproteomic identification workflow: the peptides are queried against the whole taxonomy and in the default settings, at least 150 taxa are included in the graphical model. Protein sequence homology between taxa can be expected to be strong for, e.g, housekeeping proteins,[54] but also between closely related species. Many species and their connections to peptides that are not present in the sample will therefore still be present in the graphical model, which is reflected in the high error probability β. The parameter identified for γ varies - $\gamma = 0.5$ reflects an absence of prior knowledge regarding the presence of taxa, which is often these case for metaproteomics. Lower γ can reflect the fact that, when more than 150 species are included in the graph but only 30 species are actually present in a sample, most taxa in the graphical model are more likely to be absent than present.

Reducing the number of parameter combinations searched can greatly reduce the computational demands of the Peptonizer workflow. Based on the observed parameters for the samples tested, we can make a recommendation towards the samples that should be included in the grid search for future metaproteomic analyses: for many samples $\alpha \in [0.85, 0.9, 0.99]$, $\beta \in [0.5, 0.6, 0.7]$ and $\gamma \in [0.1, 0.3, 0.5]$ should be reasonable choices and reduce the grid search space to 9 parameter combinations to be evaluated.

Memory use and runtime

The Peptonizer workflow is composed of several steps. 3 of these steps dominate the runtime and memory requirements of the workflow. These are the database search, the Unipept query and the copmutation of posteriors through belief propagation. The memory use and runtime of the database search, using X!Tandem and MS2Rescore, is largely dependent on the size of the reference database used. For the largest UNIREF50 database, the searches took more than 24h and were run with 1.8TB allocated memory.

The query time of the Unipept API endpoint depends on current Unipept server load and the number of peptides queried. Currently, a maximum of 10 peptides at once are queried in three parallel queries. If a query fails due to server overload or connectivity issues, it is retried up to two times. For about 5000 detected peptides, this can take up to several hours. From the Unipept response, the assembly of the graphical model takes a few minutes but is not optimized for memory use, requiring up to 40GB of memory.

The belief propagation runs until approximate convergence and all parameter combinations can be run fully parallelized by snakemake. The speed of convergence depends on how far the initial beliefs of the graph are from the final probability distribution. Averaged, the runtime and memory use of the previously described parameter combinations and all samples used in this study was 19min (minimum 57s, maximum 39min) for runtime and 3,36GB (minimum 1.7GB, maximum 12.9GB) for memory. We recommend running the Peptonizer in server environments.

Discussion and outlook

We have presented the Peptonizer2000, a graphical model based workflow to identify the taxonomic composition of metaprotoemic samples. The Peptonizer is currently available as a Snakemake workflow, including all analysis steps from database search to results visualization. While X!Tandem[24] is included in the workflow, any database search input can in theory be used.

The peptides provided are queried through a new Unipept API endpoint. Building a a graphical model of the probability distribution of peptides and taxa present, the Peptonizer then computes marginal probabilities of presence for all taxa included in the graph.

We have shown that the Peptonizer computes reliable probability scores for metaproteomic samples of varying complexity. In particular, we demonstrated the benefit of using Pep-

tonizer probability scores together with the number of unique peptides per taxon identified by Unipept to determine the presence or absence of specific taxa. This can especially help to justify the imposition (or not) of thresholds such as 'at least two unique peptides per taxon', a heuristic very commonly used in metaproteomics. We have also demonstrated, using a fecal sample, that the Peptonizer can identify, with reasonable probability scores, taxa (in this case, geni), that have no unique peptide detected but are included into the graphical model through shared peptides. As a drawback, one has to mention that the Peptonizer underestimated the probabilities for taxa are likely present, such as *Butyricicoccaceae*.

The analysis of more complex lab-assembled mixtures additionally showed that the Peptonizer is able to compute reliable probability assignments for most species present, but may not be able to resolve different closely related species within the same genus, as was demonstrated, e.g, by the simultaneous detection of two neighboring species from the genus *Rhizobium* with high probability scores. In this case, again, including the number of unique peptides attributed by Unipept into the decision over present or absent taxa can help resolve potential ambiguities.

Since the Peptonizer computes a probability distribution, imposing an exact threshold posterior probability for a taxon to be determined as present or absent is not justifiable. However, from the previously described results, we can estimate approximate threshold probabilities for samples of low complexity such as SIHUMIx to be above a score of 0.9, and for medium complexity samples such as the lab assembled mixture of up to 32 microorganisms, or samples analyzed with untailored but not generalist reference databases (such as SIHUMIx searched against IGC) to be around 0.8. For samples of high complexity or samples being analyzed with untailored reference databases, such as the fecal sample, we would not recommend any specific threshold value.

Currently, due to its memory use and runtime, we recommend the Peptonizer to be run in server environments. In the future, several algorithmic improvements could speed-up the

workflow. A newer version of the Unipept index is under development, substantially reducing query times.

For the computation of the taxonomic probabilities themselves, we currently use zero-lookahead belief propagation, which has been shown to be up to 5 times faster than standard belief propagation.[33] It is, however, still dependent on graph size, practically limiting the amount of taxa included in the graphical model. To address this, graph clustering approaches through community detection[55] could substantially reduce graph size, and thus computation time, by breaking up the graph of peptides and taxa into smaller, densely connected communities. Additionally, this would enable fully parallelized computation of taxon posteriors for each community.

To streamline access to the Peptonizer for researches, it will be included into the Unipept web (`https://unipept.ugent.be/`) application, enhancing the currently available taxonomic analysis based solely on unique peptides and providing a user-friendly web interface.

Lastly, the Peptonizer remains entirely dependent on the peptides detected through MS/MS experiments and downstream bioinformatic analysis. Current research improving peptide detection, such as DIA for metaproteomics[18,26] or database search engines capable to resolve chimeric spectra,[56] can be expected to translate to improved and more robust taxonomic identifications.

Conclusion

In this work, we have presented the Peptonizer2000, a graphical model based workflow for taxonomic identification of metaproteome samples. The Peptonizer2000 is based on PepGM,[27] a graphical model originally developed for strain-level identification of viral proteomes, and Unipept,[57] a pre-digested database of all Uniprot proteins and therefore peptides, with their taxonomic associations.

Different to previous methods for the analysis of metaproteomes, the Peptonizer takes peptides and their scores from a database search as input, taking into account the peptide scores themselves.

We have shown that the Peptonizer is able to compute meaningful probability scores for taxonomic identifications, providing an additional dimension to previous taxonomic assignments based solely on the number of identified peptides. The Peptonizer2000 extends the application of graphical models for taxonomic assignments of samples of unknown composition to the analysis of metaproteomes.

References

(1) Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **2015**, *348*, 1261359, Publisher: American Association for the Advancement of Science.

(2) Heyer, R.; Kohrs, F.; Reichl, U.; Benndorf, D. Metaproteomics of complex microbial communities in biogas plants. *Microbial Biotechnology* **2015**, *8*, 749–763, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1751-7915.12276.

(3) Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **2018**, *67*, 1716–1725, Publisher: BMJ Publishing Group Section: Recent advances in basic science.

(4) Kleiner, M. et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences* **2012**, *109*, E1173–E1182, Publisher: National Academy of Sciences Section: PNAS Plus.

(5) Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* **2017**, *15*, 579–590, Number: 10 Publisher: Nature Publishing Group.

(6) Pan, S.; Hullar, M. A. J.; Lai, L. A.; Peng, H.; May, D. H.; Noble, W. S.; Raftery, D.; Navarro, S. L.; Neuhouser, M. L.; Lampe, P. D.; Lampe, J. W.; Chen, R. Gut Microbial Protein Expression in Response to Dietary Patterns in a Controlled Feeding Study: A Metaproteomic Approach. *Microorganisms* **2020**, *8*, 379, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

(7) Shreiner, A. B.; Kao, J. Y.; Young, V. B. The gut microbiome in health and in disease. *Current opinion in gastroenterology* **2015**, *31*, 69–75.

(8) Ma, C.; Sun, Z.; Zeng, B.; Huang, S.; Zhao, J.; Zhang, Y.; Su, X.; Xu, J.; Wei, H.; Zhang, H. Cow-to-mouse fecal transplantations suggest intestinal microbiome as one cause of mastitis. *Microbiome* **2018**, *6*, 200.

(9) Wolf, M.; Schallert, K.; Knipper, L.; Sickmann, A.; Sczyrba, A.; Benndorf, D.; Heyer, R. Advances in the clinical use of metaproteomics. *Expert Review of Proteomics* **2023**, *20*, 71–86, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14789450.2023.2215440.

(10) Armengaud, J. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. *Environmental Microbiology* **2023**, *25*, 115–125, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.16238.

(11) Wang, Y.; Zhou, Y.; Xiao, X.; Zheng, J.; Zhou, H. Metaproteomics: A strategy to study the taxonomy and functionality of the gut microbiota. *Journal of Proteomics* **2020**, *219*, 103737.

(12) Muth, T.; Benndorf, D.; Reichl, U.; Rapp, E.; Martens, L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems* **2013**, *9*, 578–585, Publisher: The Royal Society of Chemistry.

(13) Saito, M. A. et al. Progress and Challenges in Ocean Metaproteomics and Proposed

Best Practices for Data Sharing. *Journal of Proteome Research* **2019**, *18*, 1461–1476, Publisher: American Chemical Society.

(14) Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and perspectives of metaproteomic data analysis. *Journal of Biotechnology* **2017**, *261*, 24–36.

(15) Li, L.; Ning, Z.; Cheng, K.; Zhang, X.; Simopoulos, C. M. A.; Figeys, D. iMetaLab Suite: A one-stop toolset for metaproteomics. *iMeta* **2022**, *1*, e25, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/imt2.25.

(16) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. Metaproteomic analysis using the Galaxy framework. *Proteomics* **2015**, *15*, 3553–3565.

(17) Verschaffelt, P.; Van Den Bossche, T.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept Desktop: A Faster, More Powerful Metaproteomics Results Analysis Tool. *Journal of Proteome Research* **2021**, *20*, 2005–2009, Publisher: American Chemical Society.

(18) Gómez-Varela, D.; Xian, F.; Grundtner, S.; Sondermann, J. R.; Carta, G.; Schmidt, M. Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Frontiers in Microbiology* **2023**, *14*, Publisher: Frontiers.

(19) Gouveia, D.; Pible, O.; Culotta, K.; Jouffret, V.; Geffard, O.; Chaumot, A.; Degli-Esposti, D.; Armengaud, J. Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. *npj Biofilms and Microbiomes* **2020**, *6*, 1–6, Publisher: Nature Publishing Group.

(20) Bihani, S.; Gupta, A.; Mehta, S.; Rajczewski, A. T.; Johnson, J.; Borishetty, D.; Griffin, T. J.; Srivastava, S.; Jagtap, P. D. Metaproteomic Analysis of Nasopharyngeal

Swab Samples to Identify Microbial Peptides in COVID-19 Patients. *Journal of Proteome Research* **2023**, *22*, 2608–2619, Publisher: American Chemical Society.

(21) Bender, M. A.; Farach-Colton, M.; Pemmasani, G.; Skiena, S.; Sumazin, P. Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms* **2005**, *57*, 75–94.

(22) UniProt: a hub for protein information. *Nucleic Acids Research* **2015**, *43*, D204–D212.

(23) Alves, G.; Yu, Y.-K. Robust Accurate Identification and Biomass Estimates of Microorganisms via Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2020**, *31*, 85–102.

(24) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(25) Declercq, A.; Bouwmeester, R.; Hirschler, A.; Carapito, C.; Degroeve, S.; Martens, L.; Gabriels, R. MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Molecular & Cellular Proteomics* **2022**, *21*, 100266.

(26) Pietilä, S.; Suomi, T.; Elo, L. L. Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples. *ISME Communications* **2022**, *2*, 51.

(27) Holstein, T.; Kistner, F.; Martens, L.; Muth, T. PepGM: a probabilistic graphical model for taxonomic inference of viral proteome samples with associated confidence scores. *Bioinformatics* **2023**, *39*, btad289.

(28) Mölder, F.; Jablonski, K. P.; Letcher, B.; Hall, M. B.; Tomkins-Tinch, C. H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S. O.; Kanitz, A.; Wilm, A.; Holtgrewe, M.; Rahmann, S.; Nahnsen, S.; Köster, J. Sustainable data analysis with Snakemake. *F1000Research* **2021**, *10*, 33.

(29) Joshi, C. J.; Ke, W.; Drangowska-Way, A.; O'Rourke, E. J.; Lewis, N. E. What are housekeeping genes? *PLOS Computational Biology* **2022**, *18*, e1010295, Publisher: Public Library of Science.

(30) Krause, J. L.; Schaepe, S. S.; Fritz-Wallace, K.; Engelmann, B.; Rolle-Kampczyk, U.; Kleinsteuber, S.; Schattenberg, F.; Liu, Z.; Mueller, S.; Jehmlich, N.; Von Bergen, M.; Herberth, G. Following the community development of SIHUMIx – a new intestinal in vitro model for bioreactor use. *Gut Microbes* **2020**, *11*, 1116–1129, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19490976.2019.1702431.

(31) Pfeuffer, J.; Sachsenberg, T.; Dijkstra, T. M. H.; Serang, O.; Reinert, K.; Kohlbacher, O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. *Journal of Proteome Research* **2020**, *19*, 1060–1072, Publisher: American Chemical Society.

(32) Knoll, C.; Rath, M.; Tschiatschek, S.; Pernkopf, F. Message Scheduling Methods for Belief Propagation. Machine Learning and Knowledge Discovery in Databases. Cham, 2015; pp 295–310.

(33) Sutton, C.; McCallum, A. Improved dynamic schedules for belief propagation. Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, USA, 2007; pp 376–383.

(34) Webber, W.; Moffat, A.; Zobel, J. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* **2010**, *28*, 20:1–20:38.

(35) PRIDE: The proteomics identifications database - Martens - 2005 - PROTEOMICS - Wiley Online Library. `https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200401303?casa_token=HwpGOoawKj4AAAAA%3AMjh_`

J-lycbwXZNY-R3ESEcl4-UTWNFNrkLMKt6ht5toZwcGt9ht9lzWC2-rBvgzlySrtiNe_tcKWFg.

(36) Van Den Bossche, T. et al. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nature Communications* **2021**, *12*, 7305, Number: 1 Publisher: Nature Publishing Group.

(37) Kleiner, M.; Thorson, E.; Sharp, C. E.; Dong, X.; Liu, D.; Li, C.; Strous, M. Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications* **2017**, *8*, 1558, Number: 1 Publisher: Nature Publishing Group.

(38) Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology* **2014**, *32*, 834–841, Publisher: Nature Publishing Group.

(39) Frankenfield, A. M.; Ni, J.; Ahmed, M.; Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *Journal of Proteome Research* **2022**, *21*, 2104–2113, Publisher: American Chemical Society.

(40) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926–932.

(41) Yan, P.; Sun, Y.; Luo, J.; Liu, X.; Wu, J.; Miao, Y. Integrating the serum proteomic and fecal metaproteomic to analyze the impacts of overweight/obesity on IBD: a pilot investigation. *Clinical Proteomics* **2023**, *20*, 6.

(42) Pettersen, V. K.; Dufour, A.; Arrieta, M.-C. Metaproteomic profiling of fungal gut colonization in gnotobiotic mice. *Animal Microbiome* **2022**, *4*, 14.

(43) Oren, A. Nomenclature of prokaryotic 'Candidatus' taxa: establishing order in the current chaos. *New Microbes and New Infections* **2021**, *44*, 100932.

(44) Armengaud, J.; Trapp, J.; Pible, O.; Geffard, O.; Chaumot, A.; Hartmann, E. M. Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics* **2014**, *105*, 5–18.

(45) Schiebenhoefer, H.; Van Den Bossche, T.; Fuchs, S.; Renard, B. Y.; Muth, T.; Martens, L. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics* **2019**, *16*, 375–390, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14789450.2019.1609944.

(46) Brown, M. D.; Shinn, L. M.; Reeser, G.; Browning, M.; Schwingel, A.; Khan, N. A.; Holscher, H. D. Fecal and soil microbiota composition of gardening and non-gardening families. *Scientific Reports* **2022**, *12*, 1595, Publisher: Nature Publishing Group.

(47) Almeida, A.; Nayfach, S.; Boland, M.; Strozzi, F.; Beracochea, M.; Shi, Z. J.; Pollard, K. S.; Sakharova, E.; Parks, D. H.; Hugenholtz, P.; Segata, N.; Kyrpides, N. C.; Finn, R. D. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **2021**, *39*, 105–114, Publisher: Nature Publishing Group.

(48) Jooste, P. J.; Hugo, C. J. The taxonomy, ecology and cultivation of bacterial genera belonging to the family *Flavobacteriaceae*. *International Journal of Food Microbiology* **1999**, *53*, 81–94.

(49) Ramirez Garcia, A.; Greppi, A.; Constancias, F.; Ruscheweyh, H.-J.; Gasser, J.; Hurley, K.; Sturla, S. J.; Schwab, C.; Lacroix, C. Anaerobutyricum hallii promotes the functional depletion of a food carcinogen in diverse healthy fecal microbiota. *Frontiers in Microbiomes* **2023**, *2*, Publisher: Frontiers.

(50) Zhou, J.; Wu, X.; Li, Z.; Zou, Z.; Dou, S.; Li, G.; Yan, F.; Chen, B.; Li, Y. Alterations in Gut Microbiota Are Correlated With Serum Metabolites in Patients With Insom-

nia Disorder. *Frontiers in Cellular and Infection Microbiology* **2022**, *12*, Publisher: Frontiers.

(51) Palmas, V.; Pisanu, S.; Madau, V.; Casula, E.; Deledda, A.; Cusano, R.; Uva, P.; Vascellari, S.; Loviselli, A.; Manzin, A.; Velluzzi, F. Gut microbiota markers associated with obesity and overweight in Italian adults. *Scientific Reports* **2021**, *11*, 5532.

(52) Liu, C. et al. Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **2021**, *9*, 119.

(53) Devendran, S.; Méndez-García, C.; Ridlon, J. M. Identification and characterization of a 20-HSDH from the anaerobic gut bacterium Butyricicoccus desmolans ATCC 43058 [S]. *Journal of Lipid Research* **2017**, *58*, 916–925, Publisher: Elsevier.

(54) https://orcid.org/0000 0002-9339-2511, C. J. J.; https://orcid.org/0000 0002-7047-5445, W. K.; Drangowska-Way, A.; https://orcid.org/0000 0003-0503-4181, E. J. O.; https://orcid.org/0000 0001-7700-3654, N. E. L. What are housekeeping genes? *PLoS Computational Biology* **2022**, *18*, Place: San Francisco, United States Publisher: Public Library of Science Section: Research Article.

(55) Lancichinetti, A.; Fortunato, S. Community detection algorithms: A comparative analysis. *Physical Review E* **2009**, *80*, 056117, Publisher: American Physical Society.

(56) Dorfer, V.; Maltsev, S.; Winkler, S.; Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* **2018**, *17*, 2581–2589.

(57) Verschaffelt, P.; Tanca, A.; Abbondio, M.; Van Den Bossche, T.; Moortele, T. V.; Dawyndt, P.; Martens, L.; Mesuere, B. Unipept Desktop 2.0: Construction of Targeted Reference Protein Databases for Metaproteogenomics Analyses. *Journal of Proteome Research* **2023**, *22*, 2620–2628, Publisher: American Chemical Society.