2016

An Alignment-Free "Metapeptide" Strategy for Metaproteomic Characterization of Microbiome Samples Using Shotgun Metagenomic Sequencing

Damon H. May

Emma Timmins-Schiffman

Molly P. Mikan
Old Dominion University

H. Rodger Harvey
Old Dominion University, hharvey@odu.edu

Elhanan Borenstein

See next page for additional authors

Original Publication Citation

May, D. H., Timmins-Schiffman, E., Mikan, M. P., Haryey, H. R., Borenstein, E., Nunn, B. L., & Noble, W. S. (2016). An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *Journal of Proteome Research, 15*(8), 2697-2705. doi:10.1021/acs.jproteome.6b00239

Authors
Damon H. May, Emma Timmins-Schiffman, Molly P. Mikan, H. Rodger Harvey, Elhanan Borenstein, Brook L. Nunn, and William S. Noble

Published in final edited form as:

J Proteome Res. 2016 August 5; 15(8): 2697–2705. doi:10.1021/acs.jproteome.6b00239.

Metaproteomic characterization of microbiome samples by translating shotgun metagenomic sequencing reads

Damon H. May[1], **Emma Timmins-Schiffman**[1], **Molly P. Mikan**[2], **H. Rodger Harvey**[2], **Elhanan Borenstein**[1,3,4], **Brook L. Nunn**[1], and **William S. Noble**[1,3]

[1]Department of Genome Sciences, University of Washington

[2]Department of Ocean, Earth & Atmospheric Sciences, Old Dominion University

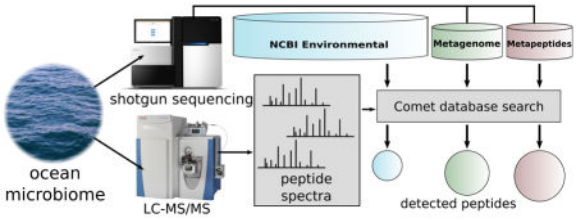
[3]Department of Computer Science and Engineering, University of Washington

[4]Santa Fe Institute

Abstract

In principle, tandem mass spectrometry can be used to detect and quantify the peptides present in a microbiome sample, enabling functional and taxonomic insight into microbiome metabolic activity. However, the phylogenetic diversity constituting a particular microbiome is often unknown, and many of the organisms present may not have assembled genomes. In ocean microbiome samples, with particularly diverse and uncultured bacterial communities, it is difficult to construct protein databases that contain the bulk of the peptides in the sample without losing detection sensitivity due to the overwhelming number of candidate peptides for each tandem mass spectrum. We describe a method for deriving “metapeptides” (short amino acid sequences that may be represented in multiple organisms) from shotgun metagenomic sequencing of microbiome samples. In two ocean microbiome samples, we constructed site-specific metapeptide databases to detect more than one and a half times as many peptides as by searching against predicted genes from an assembled metagenome, and more than three times as many peptides as by searching against the NCBI environmental proteome database. The increased peptide yield can be used to enrich the taxonomic characterization of sample metaproteomes.

Graphical abstract



Keywords

microbial ecology; metaproteomics; metagenomics; mass spectrometry; microbial communities

1 Introduction

In order to understand biogeochemical processes such as carbon cycling, ocean modelers need to develop a full understanding of the taxa that are performing important biochemical functions in the ocean. [1–3] Due to culture technique limitations on mixed microbial communities, methods for examining whole microbiomes *in situ* are needed. Metaproteomic analysis of ocean samples has the power to detect peptides from thousands of proteins over a wide range of taxonomic groups within a single analysis. [4–7] Accordingly, metaproteomics has been used to investigate the functional roles of ocean microbes in a variety of ecological contexts. [5,7,8] However, the success of high-throughput proteomics on ocean samples has been limited by a lack of detection sensitivity.

The majority of organisms active in the ocean microbiome do not have assembled genomes. [9] Public databases can provide partial metaproteome coverage, but without a precise guide to which organisms are present in the sample those databases must be extremely large in order to accommodate as much sequence variation as possible. Searching against such very large databases severely and negatively impacts search sensitivity. [10] Because of the difficulty of constructing a protein database that accurately reflects an ocean bacterial microbiome, ocean metaproteomics experiments typically only detect a small proportion of the potentially detectable peptides in a sample. [5,11]

As sequencing technologies have become more accessible, “meta-omics” studies have integrated metagenomic, metatranscriptomic and metaproteomic data. For example, databases for metaproteomic search can be constructed using genes predicted from an assembled metagenome. [8] However, this approach can lead to low peptide detection sensitivity for two reasons. First, since many gene fragments present in sequencing reads cannot be reliably assembled in to longer contigs, they will be missing from the gene prediction. Second, contig assembly necessarily involves an attempt to screen out reads with sequence errors, and the difficulty in distinguishing these from real variation present in the sample can lead to a lack of sequence coverage. For both of these reasons, at present, even metaproteomic databases based on site-specific assembled metagenomes tend to provide substantially incomplete coverage of the sample metaproteome. [12–14]

An alternative approach takes advantage of the fact that most of the organisms present in many microbiome samples are prokaryotes, and therefore high proportions of their genomes are protein-coding. Tools such as MetaGeneAnnotator, [15,16] Orphelia[13] and FragGeneScan[17] predict gene fragments directly from sequencing reads, without assembling the reads into contigs. These approaches can be used to construct metaproteomic databases that enable a greater peptide yield via database search than other methods. [14] However, the goal of these tools is sensitive gene prediction rather than peptide detection, and so databases containing translations of their raw gene fragment output can be extremely large. This can lead to impractically long running times for database searches and, more importantly, reduced peptide detection sensitivity.

In the approach described here, we begin with the gene fragments predicted by MetaGeneAnnotator or with six-frame translations of raw reads. We trim and filter these

sequences to build a database of “metapeptides”: short amino acid sequences derived from open reading frame fragments found in individual reads that are more likely to be identifiable via LC-MS/MS (Figure 1A). This approach exploits more of the metagenomic data than an approach based on an assembled metagenome, incorporating reads that fail to be integrated into a contig as well as all of the sample variation for each gene sequence, while avoiding a loss of sensitivity due to over-inclusivity. It is both more complete and more focused on the sample at hand than a strategy based on public databases, potentially including sequences never before observed in any organism and excluding sequences from species not present in the sample.

To evaluate the utility of our metapeptide approach, we compared the sets of peptide sequences detected in two Arctic Ocean microbiome samples at a 1% false discovery rate (FDR) via database search against three different databases (Figure 1B): the NCBI non-redundant database of environmental protein sequences (env_nr), which is commonly used to interrogate ocean and soil microbiome samples,[4,18] a database derived from a metagenome assembled from shotgun metagenomic sequencing of the two Arctic Ocean samples, and metapeptide databases constructed from the same sequencing reads.

Two microbiome samples were collected from the Arctic Ocean, one sample from the surface chlorophyll maximum layer in the Bering Strait (BSt) and one from bottom waters in the Chuckchi Sea (CS). A total of 1,925 peptides were detected in the BSt sample by searching the environmental database. A metagenome-derived database search yielded 2.26 times as many peptides, and a metapeptide search yielded 3.56 times as many peptides. Results were similar in the CS sample, though with many fewer peptides detected in each search. Integrating the results from all three databases further increased peptide yield.

This substantial advantage in peptide yield contributes greatly to the taxonomic classification of proteins in the samples. We used Unipept [19,20] to infer the lowest common ancestor taxon for peptides detected in each search. Comparison of the results revealed a much richer taxonomic characterization of the proteins present in the samples from the metapeptide search than from either of the other methods. Thus, in addition to dramatically increasing the number of peptides detected in a given ocean sample, the metapeptide-based approach significantly expands our understanding of the organisms producing the biochemically active molecules in a microbiome. This understanding is crucial to developing a functional model of the microbiome.

2 Methods

2.1 Experimental methods

Sample collection—Water samples were collected in August of 2013 from the Bering Strait (BSt) chlorophyll maximum layer (7m depth, 65° 43.44″ N, 168° 57.42″ W) and from the more northern Chukchi Sea (CS) bottom waters (55.5 m depth, 72° 47.624″ N, 16° 53.89″ W) using a 24-bottle CTD (conductivity, temperature and depth) rosette (10 L General Oceanics Niskin X). The integrated water column Chlorophyll-a measurement was 226.88 mg/m^2 at station BSt and 2.64 mg/m^2 at station CS. A 15-liter water sample was prefiltered through 10 μm and then 1 μm high-volume cartridges to remove larger

eukaryotes, and the filtrate comprising the bacterial microbiome was then collected on a glass fiber filter (GF/F) with nominal pore size of 0.7 μm. Filters were flash frozen and stored at −80°C until extraction.

Metagenome DNA extraction, library preparation and sequencing—Filters were sliced and DNA extraction was accomplished following methods of Wright et al. [21] Extracted DNA was sheared to < 1 kb. Excess salts were cleaned up using AMPure XP purification (Agencourt). One library for each sample was prepared using the Kapa Hyper Kit and following manufacturer's instructions (Kapa Biosystems). Library quality was assessed using Bioanalyzer before sequencing. Libraries were sequenced in one lane on an Illumina HiSeq. The resulting 100 base pair, paired-end sequencing reads were trimmed and filtered using SolexaQA,[22] with a minimum Phred quality score[23] of 20 on any base.

Protein sample preparation and tandem mass spectrometry (LC-MS/MS)—GF/F filters with the bacterial fraction were placed in 1.5 mL tubes with 100 μl of 0.5mm glass beads, 100 μl 6M urea and 500 μl nanopure water. Filters were shaken on a bead beater for one minute and then placed in ice for five minutes. This process was repeated 10 times to ensure cell lysis and filter breakup. A needle was then heated by flame and used to create a <0.5mm hole at the bottom of the 1.5mL sample tube. The sample tubes were then placed atop an open 1.5mL tube and centrifuged (3000 x g, 10 minutes). This process was completed to isolate protein lysate from extracted particles and glass beads. Protein concentrations were determined using BCA colormetric assay; 100 μg of total protein was used for digestion. Each 100 μg protein sample received 300 ng purified human ApoA1 to monitor protein digesion. Samples were reduced, alkylated, enzymatically digested with trypsin and desalted following Nunn et al. [24] Prior to MS injections, 50 fmol of the Pierce Peptide Retention Time Standard (ThermoFisher Scientific) was added to each autosample vial at 50 fmol per 2 μg total protein. Peptides were separated using an inline NanoAquity HPLC with a 4 cm pre-column (5 μm; 200A; Magic C18) and 30 cm Reprosil-Pur Basic 3 μm C18 analytical column (Dr. Maisch GmbH, Germany). Peptides were eluted using a 2–30% ACN, 0.1% formic acid non-linear gradient in 120 minutes at 300 nl/min. LC-MS/MS was performed with a Q-Exactive-HF (ThermoScientific) on technical triplicates for each sample. Instrument was operated in Top 20 data-dependent acquisition mode, collecting data on 400–1600 m/z range with a 5 s dynamic exclusion.

2.2 Computational methods

All computation was performed on a Univa Grid Engine cluster with 1.90GHz AMD Opteron processors.

Gene prediction from shotgun sequencing with existing methods—The MOCAT pipeline[25] was used to assemble a metagenome and predict genes as follows. Trimmed and filtered reads from both BSt and CS samples were aligned to the human hg19 reference using SOAPaligner v2.21, and aligned reads were removed. The remaining reads were assembled into contigs and scaftigs with SOAPdenovo v1.06. The assembly was revised, correcting for indels and chimeric regions, with SOAPdenovo v1.06 and BWA v0.7.5a-r16. Genes were predicted using Prodigal v2.60.

We used three well-established gene fragment prediction tools, MetaGeneAnnotator (in multiple species mode), FragGeneScan version 1.2.0 (illumina_10 model parameters) and Orphelia (with Net300 prediction model) to predict gene fragments directly from shotgun metagenomic sequencing reads from each sample.

Metapeptide databases—Separate metapeptide databases were constructed from the BSt and CS sequencing runs, either from predicted gene fragments or from raw read sequences. When starting from raw read sequences, each read was translated in all six reading frames, and reading frames containing a stop codon were discarded. The results described in section 3 were obtained by starting with predicted gene fragments from MetaGeneAnnotator.

Whether starting from gene fragments or from raw read sequences, amino acid sequences from each nucleotide sequence were trimmed to the first and last tryptic cleavage site (or discarded if fewer than two sites), and the remaining ends discarded (Figure 1A). This was done in order to remove partial tryptic peptide sequences that are unlikely to be detected by LC-MS/MS of a trypsinized metaproteome. The resulting candidate sequences were discarded if they were less than 10 amino acids long, if they contained no tryptic peptides with seven or more amino acids, or if the minimum Phred quality score over the length of the sequence was less than 30. Finally, metapeptide candidates meeting all the above criteria were discarded if they were represented by fewer than two reads. A FASTA database was constructed from the remaining metapeptides.

For purposes of comparison, we also made use of a metagenome-derived database of translated genes from the metagenome described above and the NCBI non-redundant database of protein sequences from large environmental sequencing projects (‘env_nr’, downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/env_nr.gz on December 1, 2015).

Database search—All database searches were performed using Comet [26] version 2015.01 rev. 2, using a concatenated decoy database in which peptide sequences were reversed but C-terminal amino acids left in place. Search parameters included a static modification for cysteine carbamidomethylation (57.021464) and a variable modification for methionine oxidation (15.9949). Enzyme specificity was trypsin, with three missed cleavages allowed. Parent ion mass tolerance was set to 50ppm around five isotopic peaks, and fragment ion binning was 0.02, with offset 0.0. Peptide-spectrum matches (PSMs) from all technical replicates were combined into a single dataset. The dataset was reduced to the highest-scoring PSM for each unique peptide sequence, irrespective of charge state, and peptide-level FDR was calculated using target and decoy peptide counts as described previously. [27]

Results of searches of individual samples against multiple databases were integrated as follows. PSMs from searches against all databases were combined into a single tab-delimited file of features for input to Percolator. [28] For each database, a new binary feature was added to the combined feature file indicating whether the PSM was derived from a search against that database. Percolator was then used to analyze the combined set, thereby computing a discriminant score for each PSM. For each scan with multiple PSMs (from

multiple databases), all but the highest-scoring PSM were removed. Peptide-level FDR was then calculated as described above.

Taxonomic inference—Detected peptides were given taxonomic assignments by Unipept version 1.1.0. For all tryptic peptides with no missed cleavages present in UniProtKB, Unipept assigns a lowest common ancestor (LCA) taxon from the NCBI Taxonomy Database, the most-granular taxon common to all organisms containing the peptide. For peptides with missed tryptic cleavages, Unipept calculates an LCA based on the LCAs associated with all completely-cleaved peptide sequences contained in the peptide.

3 Results

3.1 Gene fragment predictions from deep shotgun metagenomics sequencing are not directly usable for proteomics database search

Shotgun sequencing of the BSt and CS samples generated 171 million and 245 million reads, respectively. We evaluated three different gene prediction tools, FragGeneScan, Orphelia and MetaGeneAnnotator. None of these tools were originally developed for this high depth of coverage, nor have they been updated to accommodate high-depth sequencing. On the BSt reads, MetaGeneAnnotator ran to completion in 3.5 hours, producing 133 million fragments. Orphelia quickly exceeded 100GB of memory usage; as its output on smaller inputs was 33 times the size of the output of MetaGene, with no scoring mechanism to use for filtering, we decided not to pursue it further. After five days of running time, FragGeneScan had not yet completed, and its output on smaller inputs was 32 times the size of the output of MetaGene, so we decided not to pursue it further.

The 133 million fragments produced by MetaGeneAnnotator contained 451 million unique peptides and would take an estimated eight days to search. However, 331 million of the peptides in the database represented ragged ends of peptides terminating at the beginning or end of a metagenomic sequencing read, and did not represent a detectable tryptic peptide.

3.2 Environmental and assembled metagenome databases provide incomplete coverage of peptides in ocean samples

Next, we quantified the extent to which a public database and a metagenome-derived database could be used to detect the peptides present in the two ocean microbiome samples. All three replicates of each sample were searched against the environmental and metagenome databases, and the set of peptides detected with FDR < 0.01 in searches against each database was determined with Percolator, as described above. In the BSt sample, a metagenome-derived database search yielded 2.26 times as many peptides as an environmental database search (Figure 2), with 18.7% of all detected peptides found in both searches. In the CS sample, a metagenome-derived database search yielded 1.57 times as many peptides as an environmental database search (Figure 2), with 19.5% of all detected peptides found in both searches. The high complementarity of these two results indicates that large numbers of peptides are undetected by searches of each database alone. In the case of the metagenome-derived database, which contains only 17 million tryptic peptides, this is likely because peptides present in the sample are from microbes missing from the database.

In the case of the environmental database, low statistical power due to the large size of the database (323 million peptides) is an additional factor. This is investigated in Section 3.3.

3.3 Searching metapeptide databases increases peptide yield and enriches taxonomic characterization

Next, we evaluated the ability of our metapeptide databases to increase peptide detection sensitivity relative to the environmental and metagenome databases. The metapeptide databases constructed from the shotgun metagenomic sequencing reads from the BSt and CS samples contained 19 million and 23 million peptides, respectively. All MS/MS replicates of the BSt and CS ocean microbiome samples were searched against the metapeptide database constructed from the sample being searched, and the set of peptides detected with FDR < 0.01 was derived with Percolator. In the BSt and CS samples, the numbers of peptides detected were 1.57 and 1.98 times the number detected by searching against the metagenome-derived database, and 3.56 times and 3.10 times the number detected by searching against the environmental database, respectively (Figure 2).

To determine the reasons for this larger peptide yield, we compared the sets of peptides detected by searching the BSt spectra against the metagenomic, metapeptide, and environmental databases (Figure 3). Of the 6,850 peptides detected in the metapeptide search, 80.4% did not occur in the environmental database; by contrast, only 28.8% of the 1,925 peptides detected in the environmental search were absent from the metapeptide database. This discrepancy suggests that the metapeptide database contains more of the peptides present in the sample. Furthermore, of the 1,957 peptides present in both databases and detected in one or both searches, only 1.7% were detected in the environmental database search but not in the metapeptide database search. By contrast, 30.1% were detected in the metapeptide search but not in the environmental search. Since those peptides were present in the environmental database, we conclude that the failure to detect them is due to a loss of statistical power stemming from the much larger size of the environmental database.

By themselves, detected peptide sequences provide limited information about a sample. However, the peptides can be used to provide important insight into the sample's community composition. Accordingly, we assessed the extent to which the additional peptides detected using the metapeptide database enrich the taxonomic classification of the metaproteome. We used the Unipept tool to assign a least common ancestor (LCA) taxon to all possible peptides detected in a search of the BSt sample replicates against a given database. The metapeptide search detected 1.28 times as many peptides that were assigned LCAs as the metagenome-derived database search, and 1.87 times as many as the environmental database search. At every taxonomic rank more granular than class, the highest number of taxa were detected by the integrated search, followed by the metapeptide, metagenome and then environmental searches (Figure 4). The same order was observed when examining the number of peptides with an LCA at each taxonomic rank.

Because many metapeptides are likely from unsequenced microbes not present in public protein databases (and therefore uninformative to Unipept), an important question is whether the detected peptides that were present in the metapeptide database but absent from the environmental database conferred any taxonomic information via this method. The

proportion of detected peptides from the metapeptide database that are assignable to an LCA by Unipept is much greater for those peptides that are present in the environmental database (63.7%) than for those that are absent (27.9%). However, because 80.4% of the peptides detected in the metapeptide database search are absent from the environmental database, in absolute terms 12% more of the peptides assignable to LCAs come from that group. Thus, both the greater sensitivity and the greater coverage afforded by the metapeptide database contribute to its increased potential for metaproteome taxonomic classification.

3.4 Combining results from multiple databases further increases peptide coverage

Although the metapeptide databases are the most valuable individual databases for searching these samples, a higher overall peptide yield can be obtained by combining results from multiple databases. PSMs from searches against the environmental, metagenome and metapeptide databases were integrated as described above. In the BSt and CS samples, respectively, 1.09 times and 1.07 times as many peptides were detected by this method as by searching against the individual metapeptide databases (Figure 2). Furthermore, the integrated searches of the BSt and CS samples detected 1.13 and 1.10 times as many peptides assignable to LCAs compared with a metapeptide search (Figure 4 for BSt comparison), with more taxa observed at every taxonomic rank lower than superkingdom.

This method of integrating database search results yielded 13% more peptides at FDR < 0.01 as searching a concatenated database combining the environmental, metagenome and metapeptide databases. Due to the reduced statistical power of a search against a larger database, searching the concatenated database yielded 3.2% fewer peptides than searching the metapeptide database alone. The extra Percolator features representing the database against which each PSM was made were of modest benefit, increasing peptide yield by 1.9% *vs.* an integrated search with those features removed.

3.5 Metapeptide databases from two microbiome samples can be used to interrogate each other

Constructing a metapeptide database is a relatively expensive endeavor, requiring library preparation, short read sequencing, and computational time, and so it would be convenient to use a single database to interrogate the metaproteome from multiple samples. Our two samples are from two different locations and from two very different positions in the water column (chlorophyll maximum layer and bottom water). In each case, overall peptide yield from a database search against the metapeptide database derived from the other sample was a large improvement over the yield from a search against the environmental database (2.2 and 2.3 times as many peptides, respectively). In each case, however, searching a sample against its site-specific metapeptide database detected many more peptides than searching against the database derived from the other sample. Notably, the BSt sample appeared to benefit greatly from a search against the BSt database rather than against the CS database (1.54 times as many peptides), while the effect in the opposite direction was not as pronounced (1.42 times as many). A potential explanation for this difference lies in the depth from which the two samples were taken: the BSt sample, from the upper water column, is expected to contain more biodiversity than the CS sample taken from the bottom layer, which has no light.

3.6 Filtering protocols are critical to resulting metapeptide database size

Prior to filtering, the trimmed MetaGene output contained 120 million tryptic peptides. To investigate the effects of the filtering criteria, we systematically varied each parameter while leaving the remaining parameters set to the values described in Section 2.2. The results (Figure 5) demonstrate that filtering metapeptides based on the support of two or more reads and the use of MetaGeneAnnotator fragments rather than six-frame translations of raw reads had a particularly large effect on database size, reducing the number of unique tryptic peptides by 75% and 53%, respectively, and decreasing search time by similar proportions.

To investigate the effect of filtering parameters on database sensitivity, we generated a small sample set of 24,000 MS/MS spectra from the BSt sample (8,000 random spectra from each replicate run) to compare the number of peptides detected at FDR < 0.01 by searching each database. Beginning with MetaGeneAnnotator fragments rather than with a six-frame translation of raw reads increased detected peptides by 8.3%, demonstrating that the MetaGeneAnnotator is valuable but not crucial to the metapeptide strategy. The MetaGeneAnnotator score was not useful as a filtering criterion: higher score thresholds resulted in monotonically lower peptide yield. Requiring two or more reads increased detected peptides by 6.8%. Higher read count thresholds monotonically reduced yield. Sufficiently restrictive values for each parameter reduce peptide yield much more greatly (data not shown). However, for all parameters, within the range of values shown in Figure 5 the reduction in yield was minor, suggesting a relative robustness of the parameters.

4 Discussion

In this work, we have demonstrated the value of interrogating microbial metaproteomes by constructing metapeptide databases from site-specific shotgun metagenomic sequencing reads. These databases afford much greater peptide detection sensitivity than the NCBI environmental database or a database of genes predicted from an assembled metagenome. Furthermore, we have shown that a database derived from one sample can be used to interrogate another sample from a different location and position in the water column. By combining metapeptide databases from a variety of samples, sequencing efforts could potentially be centralized to an extent, and metapeptide databases integrated into existing metaproteomics workflows such as the MetaProteomeAnalyzer. [29] In principle, these methods should be applicable to other microbiomes, such as riverine and soil-derived microbial communities, in which prokaryotes dominate the microbiome and the great majority of organisms are unsequenced. These methods may also provide additional sensitivity in a better-understood environment such as the human gut microbiome.

From a practical standpoint, the much larger environmental database required much more computational time for database search than the metapeptide databases. On our hardware, the three BSt sample replicates, with an average of 104,000 MS/MS scans, took us an average of 11 hours to search against the BSt database and an average of 55 hours to search against the environmental database. The strategy of trimming reads to the outermost tryptic sites and the filtering criteria applied are responsible for the much smaller database size, making the metapeptide database easier to integrate into a proteomics pipeline.

De novo sequencing is another strategy for increasing peptide detection sensitivity. Although this method can yield many confident partial peptide sequences, it is less effective at confidently detecting full-length peptides. Furthermore, due to codon degeneracy, *de novo* sequencing also cannot easily link detected peptides with their corresponding nucleotide sequences for taxonomic annotation. As others have noted, the space of peptides likely to be present in a metaproteomics sample should remain tractable to a database search approach if search databases are constructed with an eye toward detection sensitivity. [30] However, *de novo* sequencing remains a viable approach for assignment of spectra that cannot be assigned with database search.

Some of the peptide sequences detected by metapeptide database search are present in organisms with publicly available genomes, enabling putative taxonomic assignment using existing peptide-based tools and enriching taxonomic characterization. However, a large proportion of peptides detected by searching against metapeptide databases have never been reported in an assembled genome. In future work, we will place those peptides within a taxonomic hierarchy using sequence homology. This will be accomplished using all of the nucleotide sequences of the reads that contributed to the inclusion of each metapeptide in the database.

Sequence homology could also be used to infer the putative molecular function of these detected peptides. With both taxonomic and functional assignments, a large number of detected peptides could be used in comparisons of the activity of various microbes between samples. This research will quantify the protein functions responsible for chemical transformations at meaningful taxonomic levels, thereby exposing microbial ecosystems at the molecular level to improve our understanding of their interactions and biological roles. Applying this approach in conjunction with recent advances in quantitative proteomics can bring about a fundamental change in how we view, analyze, and model microbial ecosystems.

An important area for future research lies in the development of improved methods for combining search results from multiple databases. The approach we have adopted here relies upon the machine learning algorithm Percolator to calibrate scores between the different database searches. A more powerful approach might be to adopt a strategy similar to cascade search, [31] searching against, in order, the metapeptide, metagenome and environmental databases. In future work, we plan to develop and validate a statistical method for combining cascade search with a machine learning post-processing step like Percolator.

The software tools described here have been implemented in Python 2.7. The software (including source code) and data described in this manuscript may be downloaded at http://noble.gs.washington.edu/proj/metapeptide.

Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number P41 GM103533, the National Science Foundation Directorate for Geosciences under award numbers OCE-1233014 and OCE-1233589, and the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

References

1. Allison SD, Martiny JBH. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(Suppl):11512–9. [PubMed: 18695234]
2. Pinhassi J, Azam F, Hemphälä J, Long RA, Martinez J, Zweifel UL, Hagström Å. Aquatic Microbial Ecology. 1999; 17:13–26.
3. Azam F, Fenchel T, Field JG, Gray JC, Meyer-Reil LA, Thingstad F. Marine Ecology Progress Series. 1983; 10:257–264.
4. Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. The ISME Journal. 2010; 4:673–685. [PubMed: 20164862]
5. Georges AA, El-Swais H, Craig SE, Li WK, Walsh DA. The ISME Journal. 2014; 8:1301–1313. [PubMed: 24401863]
6. Yoshida M, Yamamoto K, Suzuki S. Journal of Oceanography. 2013; 70:105–113.
7. Hawley AK, Brewer HM, Norbeck AD, Paša-Tolic L, Hallam SJ. Proceedings of the National Academy of Sciences. 2014; 111:11395–11400.
8. Teeling H, et al. Science. 2012; 5567:608–611.
9. Rappé MS, Giovannoni SJ. Annual review of microbiology. 2003; 57:369–94.
10. Nesvizhskii AI. Journal of Proteomics. 2010; 73:2092–2123. [PubMed: 20816881]
11. Keiblinger KM, Wilhartitz IC, Schneider T, Roschitzki B, Schmid E, Eberl L, Riedel K, Zechmeister-Boltenstern S. Soil Biology and Biochemistry. 2012; 54:14–24. [PubMed: 23125465]
12. Xiong W, Abraham PE, Li Z, Pan C, Hettich RL. Proteomics. 2015; 15:3424–3438. [PubMed: 25914197]
13. Hoff KJ, Lingner T, Meinicke P, Tech M. Nucleic Acids Research. 2009; 37:101–105.
14. Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, Pan C, Shah M, Mongodin EF, Jansson JK, Fraser-Liggett CM, Hettich RL. PLoS ONE. 2011; 6
15. Noguchi H, Park J, Takagi T. Nucleic Acids Research. 2006; 34:5623–5630. [PubMed: 17028096]
16. Noguchi H, Taniguchi T, Itoh T. DNA Research. 2008; 15:387–396. [PubMed: 18940874]
17. Rho M, Tang H, Ye Y. Nucleic Acids Research. 2010; 38:1–12. [PubMed: 19843612]
18. Moore EK, Nunn BL, Faux JF, Goodlett DR, Harvey HR. Limnology and Oceanography: Methods. 2012; 10:353–366.
19. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Journal of Proteome Research. 2012; 11:5773–5780. [PubMed: 23153116]
20. Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. Proteomics. 2015; 15:1437–1442. [PubMed: 25477242]
21. Wright JJ, Lee S, Zaikova E, Walsh Da, Hallam SJ. Journal of Visualized Experiments: JoVE. 2009:3–6.
22. Cox MP, Peterson DA, Biggs PJ. BMC Bioinformatics. 2010; 11:485. [PubMed: 20875133]
23. Ewing B, Hillier L, Wendl MC, Green P. Genome Research. 2005:175–185.
24. Nunn BL, Slattery K, Cameron Ka, Timmins-Schiffman E, Junge K. Environmental microbiology. 2014; 17:2319–2335.
25. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. PLoS ONE. 2012; 7:1–6.
26. Eng JK, Jahan TA, Hoopmann MR. Proteomics. 2012; 13:22–24. [PubMed: 23148064]
27. Granholm V, Navarro JF, Noble WS, Käll L. Journal of Proteomics. 2013; 80:123–131. [PubMed: 23268117]
28. Käll L, Canterbury J, Weston J, Noble WS, MacCoss MJ. Nature Methods. 2007; 4:923–25. [PubMed: 17952086]
29. Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, Hoffmann M, Lehtevä M, Reichl U, Martens L, Rapp E. Journal of Proteome Research. 2015 150223140604002.
30. Saito MA, Dorsk A, Post AF, Mcilvin MR, Rappé MS, Ditullio GR, Moran DM. Proteomics. 2015; 15:3521–3531. [PubMed: 26097212]

31. Kertesz-Farkas A, Keich U, Noble WS. Journal of Proteome Research. 2015; 14:3027–3038. [PubMed: 26084232]

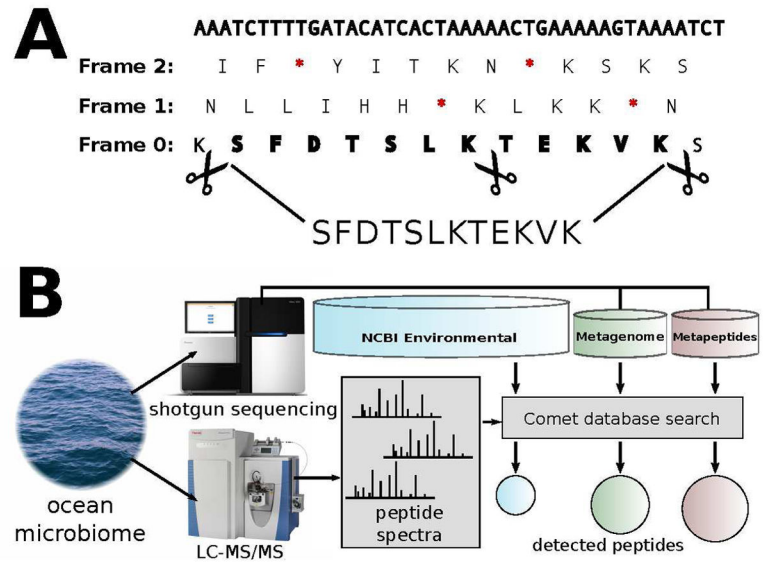


Figure 1. Multiple approaches for metaproteomics of microbiome samples

A. Metapeptide database construction begins with predicted gene fragments or with six-frame translations of raw sequencing reads. Amino acid sequences without stop codons are trimmed to their outermost tryptic sites to construct metapeptide sequences. B. Alternative proteomics workflows. Microbiome samples are subjected to shotgun metagenomic sequencing and LC-MS/MS analysis. MS/MS spectra are searched with Comet against the NCBI environmental database, against predicted genes from an assembled metagenome, or against a metapeptide database, resulting in peptide yields of different size.

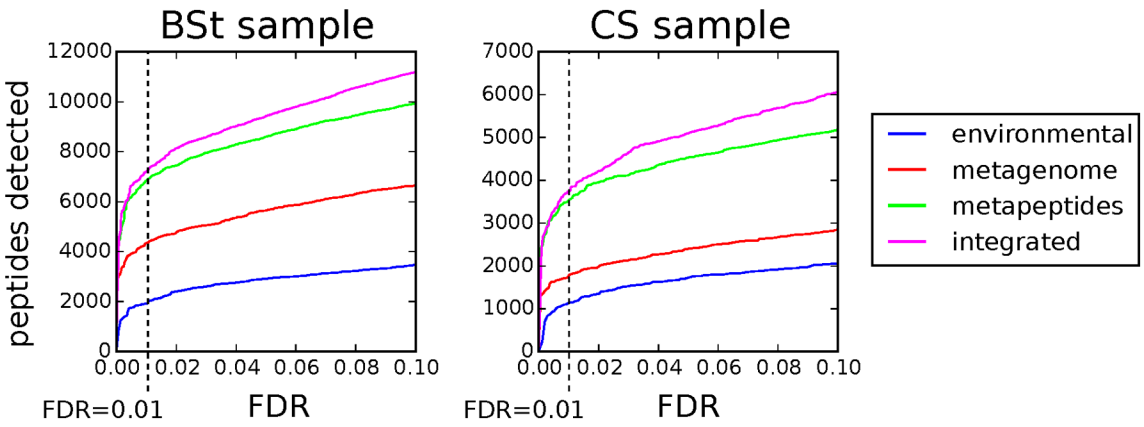


Figure 2. Peptides detected in different searches

Line plots of peptide FDR (horizontal axis) *vs.* number of peptide sequences detected at that FDR in the Bering Strait (BSt) and Chukchi Sea (CS) samples (vertical axis), when searched four different ways. Dashed line indicates peptide yield at FDR < 0.01 from searches against the NCBI environmental database (1,925 in BSt and 1,144 in CS), against the metagenome-derived database (4,352 and 1,792), against metapeptides derived from the sample being searched (6,850 and 3,544), and integrated results from all three databases (7,462 and 3,776).

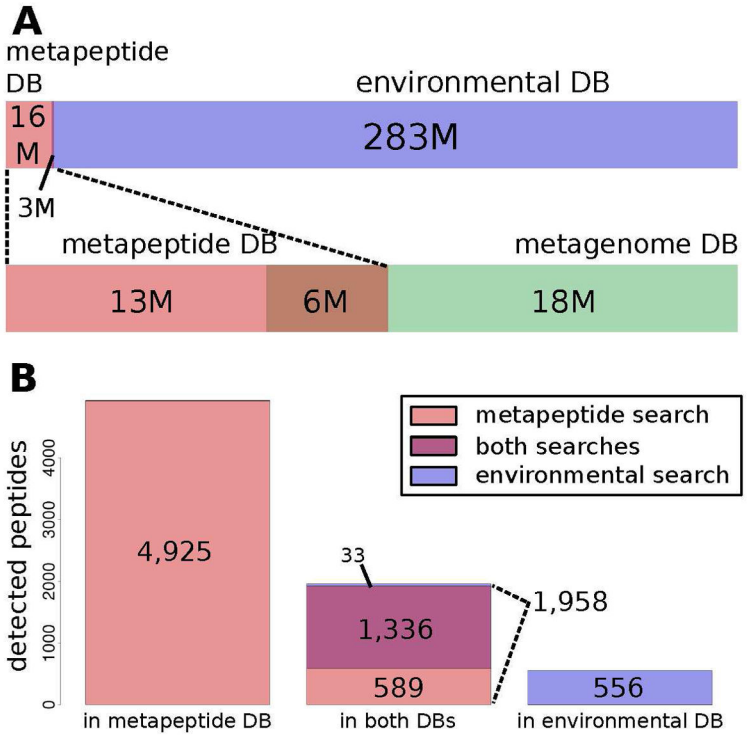


Figure 3. Database and detected peptide comparisons

A. The BSt metapeptide database contains roughly 19 million tryptic peptides. The environmental database contains 286 million, with an intersection between the two databases of 3 million peptides. The metagenome database contains 24 million peptides, with 6 million peptides in common with the BSt metapeptide database. B. Searching against the BSt metapeptide database detects 6,850 unique peptides at FDR < 0.01, *vs.* 1,925 when searching against the environmental database, with 1,336 in common. Of the 1,958 peptides detected in either search that were present in both databases, 1,336 were detected in both searches, 589 were only detected in the BSt metapeptide database search and 33 were only detected in the environmental database search.

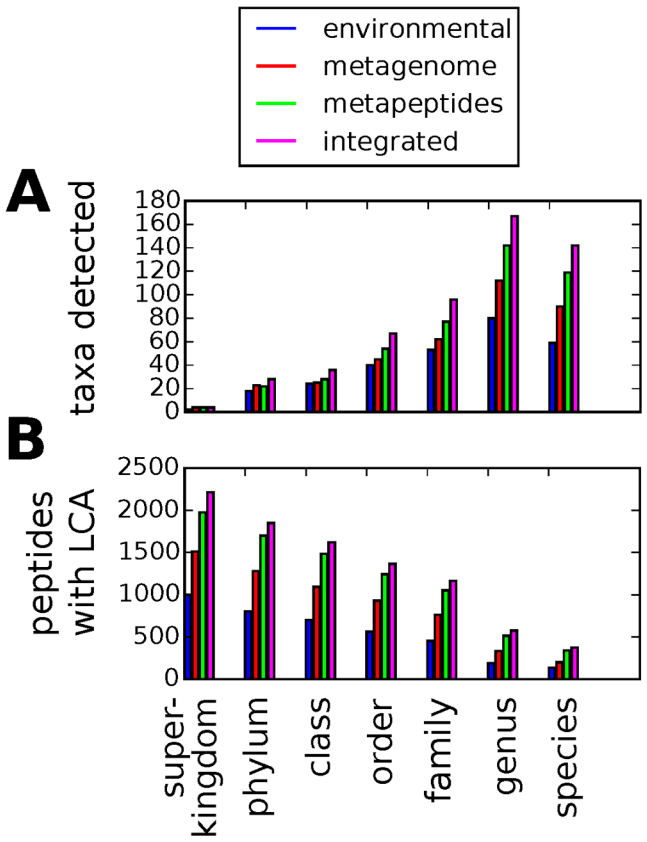


Figure 4. Taxonomic inference summary

Bar charts comparing the taxonomic information derived from four different searches of the BSt sample: against the NCBI environmental database, against the metagenome-derived database, against site-specific metapeptides, and integrating results from all three databases. A. Counts of taxa detected, by rank from superkingdom to species. B. Counts of peptides associated with an LCA at each rank.

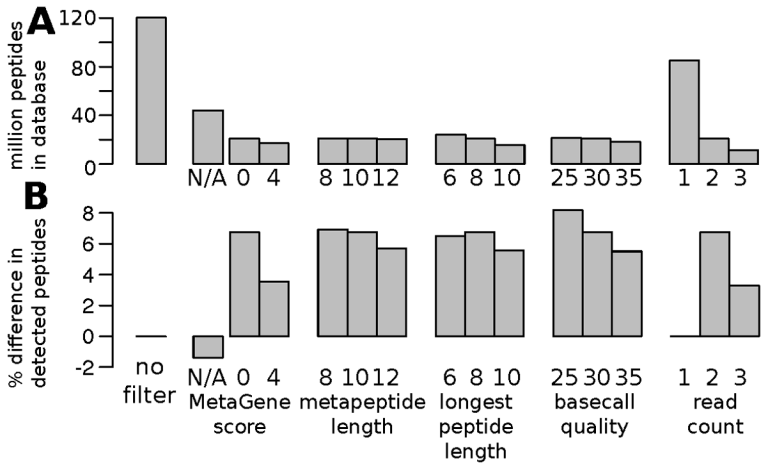


Figure 5. Metapeptide parameter comparison

Comparisons between metapeptide databases constructed with different filtering parameters. The bars on the far left represent the unfiltered, trimmed MetaGene database. Each group of three bars represents three different values for a single parameter, with all other parameters set as described in Section 2.2. In each group, the middle value represents the value used to generate the results described in Figures 2–4. The N/A value for MetaGene score represents the use of raw six-frame translations of reads instead of MetaGene output. A. Millions of tryptic peptides in each database. B. Percent difference in counts of peptides detected in a search of 24,000 scans from the three BSt sample replicates at FDR < 0.01 against each database, as compared with search against unfiltered, trimmed MetaGene database.