Detecting peptides outside the database using *de novo* sequencing and database search

William Stafford Noble[1,2] and Uri Keich[3]

[1]Department of Genome Sciences, University of Washington
[2]Paul G. Allen School of Computer Science and Engineering, University of Washington
[3]School of Mathematics and Statistics F07, University of Sydney

May 1, 2024

Most proteomics mass spectrometry analysis pipelines take as input a collection of observed mass spectra and aim to produce as output a corresponding list of peptides or proteins, along with corresponding quantifications. In this paradigm, the implicit hypothesis being tested is that "Peptide x was responsible for generating the observed spectrum y." Downstream analysis then focuses on questions such as whether the observed peptide or protein quantifications differ among experimental groups. Typically, the core of such a pipeline involves searching the observed spectra against a provided protein database, although alternative approaches that involve *de novo* sequencing or some combination of *de novo* and database search are also possible.

In a metaproteomics setting, the presence or absence of a particular peptide in the sample is not always of primary interest. When studying, for example, the microbial population in a set of samples from ocean water, soil, or the human gut, two general classes of hypotheses are common: hypotheses regarding protein functional dynamics in the microbial population and questions about the distribution of taxonomic groups in the microbial population. In either of these cases, we argue, the primary hypothesis should not be framed as "Peptide x was responsible for generating the observed spectrum y." Instead, the question is whether "The peptide that was responsible for generating the observed spectrum y is identical or homologous to peptide x."

To understand the motivation for this shift in hypotheses, consider the setting in which a metaproteomics sample is subjected to *de novo* sequencing analysis. In practice, the list of peptides that are produced by this analysis may be used to test hypotheses about the associated functional annotations, as represented, for example by KEGG or GO annotations. Alternatively, that same list may be subjected to taxonomic analysis by mapping the peptides onto taxonomic groups. In either of these settings, a truly novel peptide that has never been seen before is not of much use. Such a peptide cannot be associated with a protein ID and hence cannot be mapped to a functional label or a taxonomic group. On the other hand, if that novel peptide p can be mapped onto a known peptide p_n using a sequence similarity tool such as BLAST [1], then it may still be possible to use functional or taxonomic annotations associated with p_n [7, 5, 8].

Insert paragraph about related work.

In this work, we propose a two-pronged analysis strategy that subjects a given collection of metaproteomics data to both a database search and *de novo* sequencing, with the aim of producing a list of peptides that are identical or homologous to peptides in the sample (Figure 1). The procedure involves merging the peptides detected by the two methods into a single list and then searching that list against a protein database using a variant of the Smith-Waterman sequence alignment algorithm. Finally, the peptides resulting from this similarity search are subjected to a rigorous false discovery rate (FDR) control procedure, called RESET [3], which makes use of scores from the database search, the *de novo* sequencing, and the Smith-Waterman search. Critical to the success of this approach is the inclusion of randomized decoys in the protein database, which enable RESET to accurately estimate the FDR in the combined list.

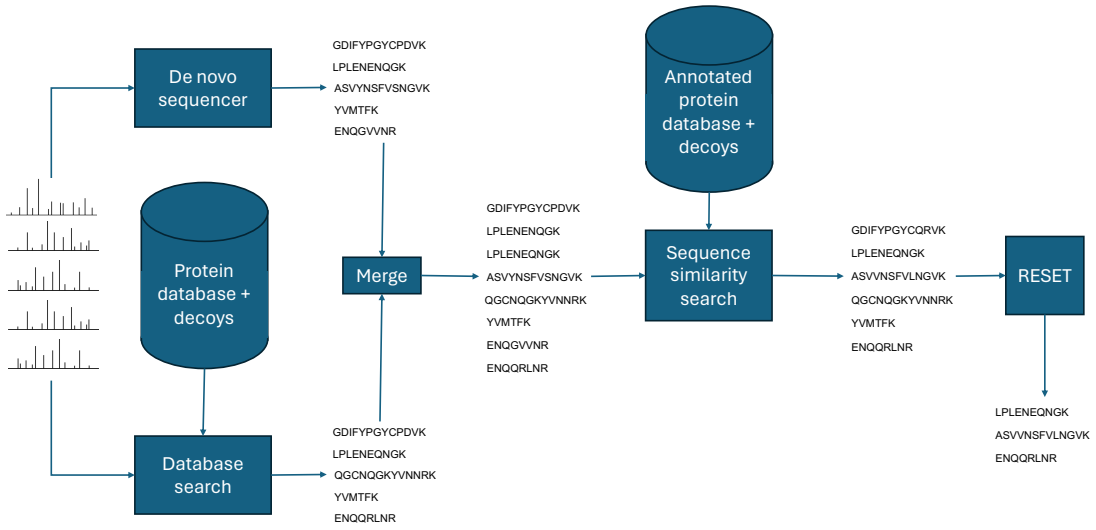

Figure 1: **Two-pronged analysis strategy.**

1 Methods

1.1 The denovo-db protocol

The denovo-db[1] procedure involves four steps (Figure 1), which we describe in detail here.

1.1.1 Database search

The database search step is carried out using Comet [2], using a database containing environmental proteins (specify) as well as annotated proteins (UniProtKB version XXX). Rather than using Comet's built-in functionality to generate decoys, which involves independently reversing each target peptide, we generate decoy proteins in advance by reversing each protein sequence. A concatenated target+decoy fasta file is then provided as input to Comet. We run Comet using semi-tryptic digestion without proline suppression of cleavage, allowing for up to three missed cleavages per peptide. We include one static modification, carbamidomethylation of cysteine (C+57.021464), and one variable modification, oxidation of methionine (M+15.9949, maximum three per peptide). The precursor m/z window is set to 10 ppm, and we allow up to three isotope errors. The fragment m/z axis is discretized with bins of 0.02 m/z. All other Comet parameters are set to their default settings. We retain a single, top-scoring target or decoy peptide-spectrum match (PSM) for each spectrum. We then reduce this set of PSMs to a list of distinct peptide sequences, with modifications removed, retaining only the maximum score per peptide.

1.1.2 *De novo* search

For *de novo* sequencing, we use Casanovo version 4.1.0 [9]. We run the Casanovo "sequence" command using default settings. Casanovo uses a fixed set of seven types of variable modifications (methionine oxidation, asparagine deamidation, glutamine deamidation, N-terminal acetylation, N-terminal carbamylation, N-terminal NH_3 loss, and the combination of N-terminal carbamylation and NH_3 loss). We effectively disable Casanovo's precursor m/z filtering by adding 1 to any negative Casanovo score, thereby ensuring that all scores are in the range $[0, 1]$. As in the database search setting, we first retain a single top-scoring PSM for each spectrum and then collapse the list to the maximum score per distinct, unmodified peptide.

[1] Placeholder name

Category	Feature	Type
glsearch	Database peptide length	integer
glsearch	Sequence similarity score	real
glsearch	Percent identity	real
Casanovo	Number of matched spectra	integer
Casanovo	Maximum score	real
Casanovo	Charge states	multiple Boolean
Casanovo	Delta m/z	real
Casanovo	PTMs	multiple Boolean
Comet	Number of matched spectra	integer
Comet	Tryptic cleavage at N-terminus	Boolean
Comet	Tryptic cleavage at C-terminus	Boolean
Comet	Maximum score	real
Comet	Charge states	multiple Boolean
Comet	Delta m/z	real
Comet	PTMs	multiple Boolean

Table 1: **Features input to RESET.** Each entry corresponds to a database peptide and its associated query peptide. The query peptide may have been detected in the database search, *de novo* sequencing, or both. The delta m/z values are in units of ppm, and are adjusted to account for isotope errors. Possible additional features: glsearch percent identity, Booleanized delta m/z features.

1.1.3 Glsearch

The lists of peptides nominally detected by Comet and Casanovo are merged into a single list, and this list is searched against the target+decoy database using the "glsearch36" command from the FASTA package, version 36.3.8 (`https://fasta.bioch.virginia.edu/fasta_www2/fasta_intro.shtml`) [6]. This command implements a global-vs-local ("glocal") version of the classic Smith-Waterman dynamic programming algorithm, in which the query sequence is matched against a database protein while requiring that the complete query be included in the match but allowing only a portion of the database protein to be included in the match. We set the gap open penalty to 0, the gap extension penalty to 33, and we use the BLOSUM62 substitution matrix. For each query, we retain the top-scoring database peptide (or peptides, in the case of ties), along with the associated protein IDs.

1.1.4 RESET

In the final step of denovo-db, the database peptides output by glsearch are provided as input to the RESET algorithm [3], each coupled with a feature vector that is used by RESET to carry out FDR control. Each of the database peptides output by glsearch is matched to a query peptide that was produced by the Comet search, by Casanovo, or by both. Accordingly, the feature vector contains three different categories of features: those that are properties of the database peptide, the Comet peptide, and the Casanovo peptide. For database peptides that do not map to either a Casanovo peptide or a Comet peptide, the corresponding scores are all set to 0. RESET uses the given feature vectors to train a machine learning model to rank the peptides and uses the decoys to set an FDR threshold in the ranked list, as previously described [3].

1.2 Data

N.B. I am copying all this info here for reference. Will reduce to relevant bits later. —WSN

1.3 Sample collection and water column analysis

Sample collection methods are detailed in Nunn et. al., 2024 (in review). Briefly, field sampling was conducted in East Sound, WA from May 27 - June 18, 2021 by pumping water from 100 m offshore at a constant depth of 2 m and flow rate of 5 L min-1. Whole water was filtered through a 200 µm nylon mesh followed by

three subsequent filters at 100 µm, 10 µm, 1 µm before final collection onto a 0.22 µm polyethersulfone (PES) 47 micron filter. Final metaproteomic samples were stored long term in -80°C on 0.22 µm polyethersulfone (PES) filters. Additional samples collected included nutrients, flow cytometry, prokaryotic DNA/proteomics, eukaryotic DNA/RNA/proteomics, and dissolved metabolites. Water chemistry was also monitored at the collection site with a YSI Exo1 Sonde probe capturing chlorophyll a concentration, salinity, temperature, dissolved oxygen, and pH every 10 minutes.

1.4 Metagenome assembly and database generation

DNA extraction on samples from all time points collected was performed according to methods detailed in Nunn et. al., 2024 (in review). Equal amounts of DNA (50 ng) were taken from all 132 samples to generate a pooled sample for whole genome sequencing (WGS) using 150 bp paired-end Illumina sequencing (NextSeq 500 Mid Output v2 Kit - 300 cycle) through the Northwest Genomics Center (NWGC) at the University of Washington. DNA library preparation, quality controls, and WGS were performed by the NWGC. The resulting fastqs were assembled into a metagenome using MEGAHIT. Samtools 1.14 was used with Python 3.7.7 to filter only mapped reads and Prodigal 2.6.3 was used to convert contigs into a FASTA file consisting of 3,457,818 proteins (Github).

Note that individual metagenomes available through JGI - Brooks 2024 paper in review

1.5 Proteomic sample processing

Samples collected for microbiome proteomic analysis (0.22 PES) were processed using a mechanical lysis in 100 μl S-Trap solubilization buffer (5% SDS, 50 mM TEAB, 2 mM MgCl2, 1X HALT protease and phosphatase inhibitors in nanopure water) followed by three subsequent rinses of the filter with 100 μl nanopure water. The resulting 400 μl whole cell solution was collected in microfuge tubes and sonicated (Branson 250 Sonifier; 20 kHz, 30 $\times$ 10 s on ice). Samples were then evaporated using a SpeedVac to reduce the volume to 100 μL (5% SDS). Protein amounts in each sample were quantified using a Bicinchoninic Acid protein assay kit (Pierce Thermo Scientific). Enolase protein standard (0.16 μL 100 ng μl–1 enolase per 1 μg protein) was added to the sample to track digestion efficiency. The sample (20 μg protein) was treated with benzonase to degrade DNA (0.5 μL 250 unit μl–1 10 minutes at 95°C), and proteins were reduced with 20 mM dithiothreitol for 10 minutes at 60°C and 5 minute cool down to room temperature, alkylated with 40 mM iodoacetamide for 30 minutes in the dark, acidified to pH ¡2 (1.2% aqueous phosphoric acid), and then processed on an S-trap column according to manufacturer's recommendations (Protifi ref). Proteins were digested with Promega modified trypsin (2 μg for 1:10 ratio, 1 hour 37°C). Resulting peptides were evaporated to dryness and resuspended in 2% acetonitrile (ACN), 0.1% formic acid with final concentration of 0.5 μg protein μl–1. Prior to analysis on the mass spectrometer, each sample received 50 fmol ug-1 protein of Pierce Retention Time Calibration (PRTC) as an external standard to monitor MS performance across experiments.

1.6 Proteomic mass spectrometry overview

All mass spectrometry analyses were performed on an Orbitrap Lumos Fusion Tribrid instrument with an inline EASY-nLC 1200 (ThermoFisher Scientific). Reverse phase chromatography was achieved using a manufactured PicoTip fused silica capillary column (75 μm i.d., 40 cm long) packed in-house with C18 beads (Dr. Maisch ReproSil-Pur; C18-Aq, 120 Å, 3 μm) and maintained at 50°C. The analytical column was fitted with a 4 cm long, 150 μm i.d. in-house packed kasil-frit precolumn (Dr. Maisch ReproSil-Pur; C18-Aq, 120 Å, 3 μm). Peptides were eluted using an acidified (formic acid, 0.1% v/v) water-acetonitrile gradient (2-45% acetonitrile). Sample analyses on the MS were randomized to reduce batch effects and quality control (QC) peptide mixtures were analyzed every sixth injection to monitor chromatography and MS sensitivity. Each sample was analyzed with data-dependent acquisition DDA in random order and Skyline was used to evaluate the consistency of the peak areas and retention times of QC standards through all analyses. The DDA mass spectrometry data are deposited to the Proteome Xchange Consortium via the PRIDE partner repository with the dataset identifier PXD-XXXXX (username: XXX.ac.uk; password: XXX).

1.7 DDA analyses and interpretation

Samples from date 6/12/21 (17:00) to 6/18/21 (13:00) were analyzed on the mass spectrometer operating in DDA mode. The top 20 most intense ions were selected for MS2 acquisition from precursor ion scans of 400–1200 m/z. Centroid full MS resolution data was collected at 70,000 with AGC target of 1 ×106 and centroid MS2 data was collected at resolution of 35,000 with AGC target of 5 × 104. Dynamic exclusion was set to 10 seconds and +2, +3, +4 ions were selected for MS2 using DDA mode.

Comet 2022.01, was used to search the DDA files; parameters included: reverse concatenated sequence database search, 10 ppm precursor mass tolerance, semi-tryptic specificity with 3 missed cleavages allowed, cysteine modification of 57 Da (resulting from the iodoacetamide) and modifications on methionine of 15.999 Da (oxidation). Percolator (version 3.02), was used to select peptides with 1% FDR. A Nextflow workflow was used to filter the metagenomic FASTA into a smaller FASTA for proteomic interpretation by applying a q-value filter threshold to the DDA search Comet and Percolator results that returns only peptides with a q-value greater than or equal to 0.01 (Supplemental file DDA-refined FASTA). The resulting DDA-refined database contains 36,429 protein sequences (Supplemental file DDA-refined FASTA).

1.8 Competing methods

We compare denovo-db to two alternative methods for producing a list of annotated peptides from a metaproteomics dataset.

The first method directly maps the observed spectra to peptides in UnitProtKB. For this step, we use the same Comet database search procedure as in denovo-db (Section 1.1.2), applied only to the annotated (UnitProtKB) database. This search is followed by post-processing with Percolator [4] version X.XX using default parameters, producing a list of peptides subject to Percolator's 1% peptide-level FDR control. Modifications are removed from this list, and duplicates are removed.

The second method is modeled after the approach implemented in MetaGOmics. First, we use Comet to search the metaproteomics data against the full database that was used in the first step of denovo-db. This search is followed by Percolator post-processing, thereby producing a list of peptides subject to 1% FDR control, as above. This list is split into two groups: peptides that appear in UnitProtKB and those that do not. The latter set of peptides is searched against UnitProtKB using NCBI BLAST [1], with an E-value cutoff of 10^{-10} and retaining only the top hit. The peptides output by BLAST are then added to the list of UnitProtKB peptides from the database search step.

2 Results

3 Discussion

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] J. K. Eng, T. A. Jahan, and M. R. Hoopmann. Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics*, 13(1):22–24, 2012.

[3] J. Freestone, L. Käll, W. S. Noble, and U. Keich. Semi-supervised learning while controlling the fdr with an application to tandem mass spectrometry analysis. *bioRxiv*, 2023.

[4] L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–25, 2007.

[5] Hugo Kleikamp, Ramon van der Zwaan, Ramon van Valderen, Jitske van Ede, Mario Pronk, Pim Schaasberg, Maximilienne Allaart, Mark van Loosdrecht, and Martin Pabst. NovoLign: metaproteomics by sequence alignment. *bioRxiv*, pages 2024–04, 2024.

[6] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[7] Michael Riffle, Damon H May, Emma Timmins-Schiffman, Molly P Mikan, Daniel Jaschob, William Stafford Noble, and Brook L Nunn. MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes*, 6(1):2, 2017.

[8] Caitlin MA Simopoulos, Zhibin Ning, Xu Zhang, Leyuan Li, Krystal Walker, Mathieu Lavallée-Adam, and Daniel Figeys. pepFunk: a tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies. *Bioinformatics*, 36(14):4171–4179, 2020.

[9] M. Yilmaz, W. E. Fondrie, W. Bittremieux, S. Oh, and W. S. Noble. *De novo* mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the International Conference on Machine Learning*, pages 25514–25522, 2022.