Metaproteomics analysis of ground-truth dataset using predicted deep learning library searching

Pratik D. Jagtap[1], Subina Mehta[1], Andrew T. Rajczewski[1], James Johnson[1], Reid Wagner[1], Mathias Wilhelm[2], Manuel Kleiner[3], Brian C Searle[4]; Timothy J. Griffin[1]

1. University of Minnesota, Minneapolis, MN; 2. Computational Mass Spectrometry, Technical University of Munich, Freising, Germany; 3. North Carolina State University, Raleigh, NC; 4. The Ohio State University, Columbus, OH

INTRODUCTION

Mass spectrometry-based metaproteomics, which characterizes the expressed microbial proteins within a complex ecosystem, has emerged as an insightful method to unravel the mechanisms underlying microbiome dynamics. Metaproteomics data analysis presents challenges, including large protein sequence database searches which can lead to low sensitivity and/or detection of a high proportion of false positives. The emergence of deep learning-based predicted spectral library searching methods offer an opportunity to improve the detection sensitivity. In this study, we use a ground-truth dataset to test the capabilities of deep learning-based spectral library searching.

THE DATASET

A ground-truth dataset of digested mixture of 32 microbial species and strains of Archaea, Bacteria, Eukaryotes and Bacteriophages with known species abundances was used (Kleiner M *et al* (2017)). Some of the bacterial strains were very closely related, but still distinguishable at the protein and nucleotide sequence level. The uneven mock community was designed to cover a large range of species abundances both at the level of cell number and proteinaceous biomass to test for the dynamic range and detection limits of the quantification methods.

THE DATABASE

The ground-truth dataset was searched against:

1X database: protein FASTA file comprising of 112,580 protein sequences from the 32 organisms.

Protein sequences from the integrated gene catalog (IGC) were randomly sampled and added to increase database size :

2X database: 1X database + 112,580 protein sequences from IGC database.

4X database: 1X database + 337,740 protein sequences from IGC database.

10X database: 1X database + 1,014,220 protein sequences from IGC database.

SEARCH ALGORITHMS

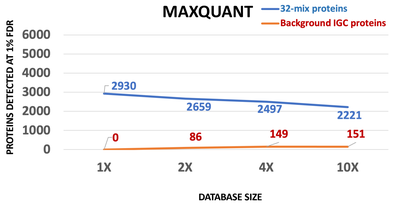
MaxQuant (version 1.6.2.3 ; Cox and Mann, 2008), FragPipe (version 19.1; Kong et al 2017) were used for protein sequences FASTA file searches while Scribe (version 2.12.30 ; Searle et al 2023) was used for spectral library search. For spectral library search Prosit (Gessulat et al 2019) was used to generate the 1X, 2X, 4X and 10X database using predicted deep learning method.

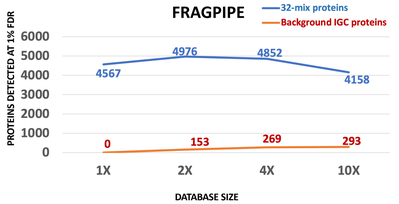
The three search algorithms were used for peptide and protein identification from four biological replicate DDA-MS datasets. The search algorithms were also used for protein and peptide quantification.

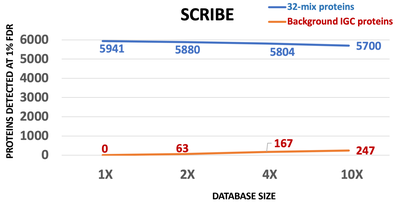
Organism abundance was calculated for high-abundance organisms, Intermediate abundance and low abundance based on protein abundance levels for these organisms. Quantitative measurements from the three search algorithms was compared with known values (Reference%) for the four database searches.

PROTEIN IDENTIFICATION

The ground-truth dataset was searched against 1X (original database), 2X (original database + 1X IGC database), 4X (original database + 3X IGC database) and 10X (original database + 9X IGC database) databases using MaxQuant, FragPipe and Scribe search algorithms.

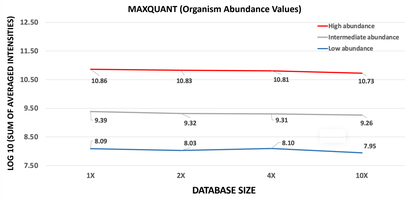


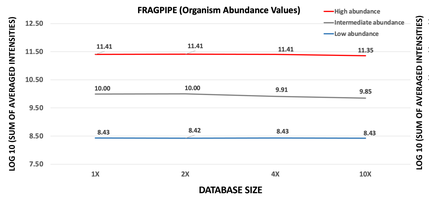


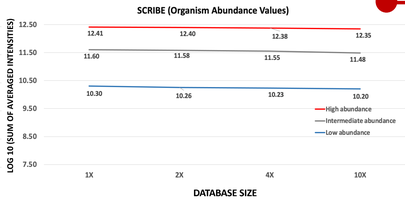


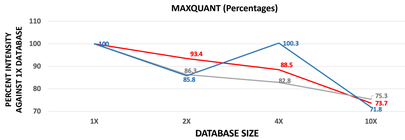
ORGANISM ABUNDANCE

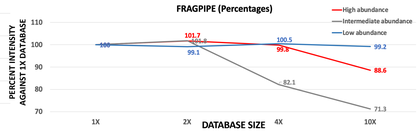
Organism abundance was calculated for high-abundance organisms, intermediate abundance and low-abundance based on protein abundance levels for the organisms.

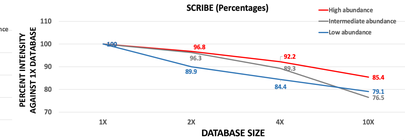






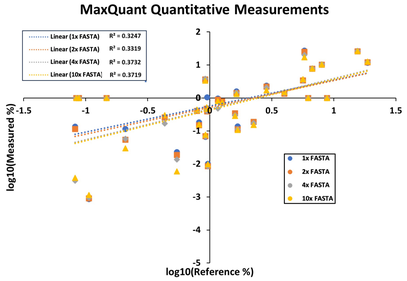


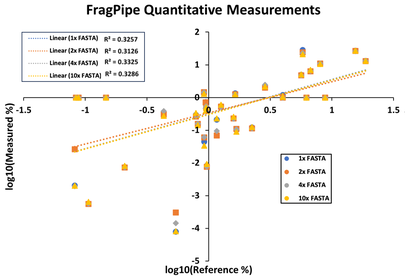


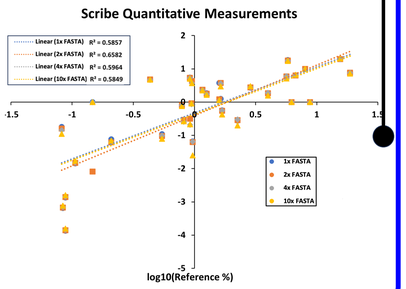


QUANTITATIVE MEASUREMENTS

Quantitative measurements from the three search algorithms was compared with known values (Reference%) for the four database searches.







RESULTS

Detection of proteins from 32-microbe mix sample dataset decreases and detection of background IGC proteins increases as the database size increases. This effect is more pronounced for protein FASTA database searching algorithms (MaxQuant and FragPipe).

Organism abundance detection decreases as database size increases for all levels of organism abundance. Scribe shows consistent decrease for all organism abundance levels.

Only Scribe detected low-abundance phage proteins such as M13 phage, ES18 phage, F0 phage and F2 phage.

Quantitative analysis of the DDA-MS data using peptide intensity values showed that none of the algorithms performed well. We plan to investigate this further to determine an appropriate method of quantitative estimation.

REFERENCES

1: Kleiner M, *et al.* (2017) Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun. ;8(1):1558.

2: Gessulat S *et al.* (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 16(6):509-518.

3: Cox J, Mann M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. ;26(12):1367-72.

4: Kong AT, *et al* (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods.; 14(5):513-520.

5: Searle BC *et al* (2023) Scribe: Next Generation Library Searching for DDA Experiments. J Proteome ;22(2):482-490.