INTRODUCTION

In metaproteomics, the choice of protein sequence search database plays critical role for identification of peptides/proteins from mass spectrometry data. Database size and composition presents challenges for optimal identification. In this study, we evaluate contemporary database generation software tools used to generate customized, compact and annotated protein sequence search databases. In particular, we test different approaches of generating protein sequence FASTA database from metatranscriptomics and metagenomics data.

- Approaches used:
  - Graph2Pro (metatranscriptomics) - a graph-centric approach that employs the de Bruijn graph structure reported by metagenome assembly algorithms to generate a comprehensive database. The assembled graph/potective peptides are compared with the supporting MS/MS spectra to obtain the potential protein sequences.
  - Sixgill (metagenomics) – This method generates protein fragments that are obtained by six frame translation of site-specific metaproteids.
  - Metagenomic binning (metagenomics) – This method involves assembling metagenomic reads in contigs using metaSPAdes, the taxonomic binning of contigs was done using MaxBin2 with a minimum contig length of 5000.
  - Sectioning of the Combined Database (Sixgill + Graph2Pro + Metagenomic Binning)

Results were compared with ComPIL 2.0 (for search datasets with unknown composition). Taxonomic and Functional annotation was performed to evaluate the outputs.

METHOD

- Metatranscriptomics / metagenomics data was obtained from a time course study of an anaerobic cellulose degrading community.
- Mass Spectrometry raw files were searched against the databases generated using Sixgill. Metagenomic binning, Graph2Pro and the combined database using SearchGUI/PeptideShaker and the sectioning method (for combined database) within the Galaxy platform.
- The PSM (1% FDR), Peptide (1% FDR) outputs generated were then compared with ComPIL 2.0 (1% FDR).
  - For functional and taxonomic annotation of these peptides, we used Uniprot 4.5 web interface.

RESULTS

Our primary evaluation shows that the searches with the sectioned combined database provides better identification statistics (such as peptides identified; taxonomy and functional assignments) as compared to other individual approaches.

- Peptide overlap between the methods.
- ComPIL 2.0 and Metagenomic binning had lesser peptide identification compared to the others, presumably due to the size of their databases.
- Combined database with sectioning approach provided with maximum number of peptide identifications.

CONCLUSION

This study evaluates the performance of different database generation methods for metaproteomics analysis. The six different approaches were compared based on their ability to identify peptides and assign taxonomic and functional annotations. The combined database approach showed the best overall performance, indicating the importance of integrating various methods for comprehensive metaproteomics analysis.