Introduction to Metaproteomics

https://gcc2022.sched.com/event/xaCl/introduction-to-metaproteomics

Galaxy Community Conference
July 17-20, 2022

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
Workshop instructors & acknowledgements

• Instructors
  • Tim Griffin
  • Pratik Jagtap
  • Subina Mehta
  • James Johnson

• Other contributors
  • Andrew Rajczewski
  • Reid Wagner
  • Katherine Do
  • usegalaxy.eu team
  • Galaxy community

https://gcc2022.sched.com/event/xaCl/introduction-to-metaproteomics
AGENDA AND SCHEDULE

1:20 PM – 1:30 PM: Introduction to Mass Spectrometry

1:30 PM - 1:50 PM: Multi-omics and Galaxy Workflows for Metaproteomics Research

1:50 PM - 3:10 PM: Taxonomy and functions of microbiome from environmental samples

3:10 PM - 3:50 PM: Summary and Questions and Answers

Please ask questions as they come up!
Workshop Topics

Mass spectrometry-based proteomic informatics

Metaproteomics (multi-omics)

Protein sequence analysis
Shotgun proteomics: breaking proteins into peptides

1. Protein, MW = 10,000 +
   
   digest into peptides (enzymatic, chemical)

2. Peptides, MW < 4,000

   Trypsin
   N-THK,NCPHIPVVGTPGR,IPD-C
   cleaves c-terminal side of arginine (R) and lysine (K)

3. Peptide fragments
Mass Spectrometry-based proteomics: A primer

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)
Matching amino acid sequences to MS/MS data

Raw MS/MS spectrum

Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification
Inferring protein identity for peptide sequences

Cytochrome C

\[ \text{NH}_2\text{GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRTGQAPGFTYTDANKNKGITW} \]
\[ \text{KEETLMEYLENPKKYIPGKMIFAGIKKTEREDLIAYLKKATNECOOH} \]
Microbiome in Numbers

- **100 Trillion** symbiotic microbes live in and on every person and make up the human microbiota.
- **95%** of our microbiota is located in the GI tract.
- **150:1** The human body has more microbes than there are stars in the milky way.
- **90%** The microbiome is more medically accessible and manipulable than the human genome.
- **5:1** The number of viruses in the gut microbiota is 5 times the number of bacteria.
- **1.3X** You have 1.3 times more microbes than human cells.
- **>10,000** The number of different microbial species that researchers have identified living in and on the human body.
- **2kg** The gut microbiota can weigh up to 2kg.
- **2.5** Each individual has a unique gut microbiota, as personal as a fingerprint.

More information can be found at:

- [https://worldmicrobiomeday.com/resources/](https://worldmicrobiomeday.com/resources/)
- [https://www.nature.com/articles/d41586-020-00193-3](https://www.nature.com/articles/d41586-020-00193-3)
Potential to unravel the mechanistic details of microbial interactions with host / environment.
**Metaproteomics**

**Bond and Wilmes 2004**

“The large-scale characterization of the entire protein complement of environmental microbiota at a given point in time”

*Environ. Microbiol. 6, 911–920.*

**Bond and Wilmes 2015**

“Through the application of metaproteomics to different microbial consortia over the past decade, we have learnt much about key functional traits in the various environmental settings where they occur.”

## Metaproteomic Challenges

### Single-Organism Proteomics
- Size: Small to medium size (10K to 100K sequences)
- Complexity: Single + contaminants

### Metaproteomics
- Size: Large (1 million and above)
- Complexity: Multi-organism database with homologous proteins

#### Search Database
- **Size:** Small to medium size (10K to 100K sequences)
- **Complexity:** Single + contaminants

- **Search Algorithms** being developed to address large and complex database searches
- **Protein Grouping** at multi-organism level
- **Identification Statistics** affected by large databases
- **Taxonomy** based on unique peptide identifications
- **Functional Analysis** based on proteins identified

Disparate tools and multiple processing steps.
The importance of the protein sequence database

- Microbe sequences
- Human sequences
- Common contaminants
Sample Collection
Water samples were collected from the Bering Strait and Chukchi Sea and oceanic marine bacteria retained on a 0.7 μM filter.

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

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

1Department of Genome Sciences and 2Department of Computer Science and Engineering, University of Washington, Seattle, Washington, 98195-5005, United States
3Department of Ocean, Earth & Atmospheric Sciences, Old Dominion University, Norfolk, Virginia 23529, United States
4Santa Fe Institute, Santa Fe, New Mexico 87501, United States

http://noble.gs.washington.edu/proj/metapeptide/
Metaproteomics Workflow

**Database Generation**
- FASTQ
- FASTA
- Mass Spectrometry Data

**Database Search & Strategies**
- Search Algorithm
- Spectra

**Functional Analysis**
- Known Function
- Peptides
- Proteins

**Taxonomy Analysis**
- Unique
- Shared
- Unassigned
Software tools can be used in a sequential manner to generate analytical workflows that can be reused, shared and creatively modified.
Hands On Session

https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaproteomics/tutorial.html

https://usegalaxy.eu

- Download and start workflow
- Observe outputs
- Tool Basics
SearchGUI matches MS/MS spectra to peptide sequences

- SearchGUI allows for multiple search engines to run simultaneously
- Specific digestion conditions can be selected
- Mass spectrometer parameters can be selected to maximize the efficacy of spectral matches
- Post-Translational Modifications (PTMs) can be added to the search parameters
PeptideShaker filters SearchGUI results.
- Search GUI results are filtered by FDR to yield most confident peptide spectral matches (PSMs)
- Peptide Shaker generates outputs such as Protein Report, Peptide Report and mzIdentML files for subsequent analysis.

**FUNCTIONAL ANALYSIS**

Unipept

**TAXONOMY ANALYSIS**

Peptides → Proteins

- Known Function
- Unassigned
- Unique
- Shared

Taxonomic and Functional analysis with Unipept

https://unipept.ugent.be/publications
Who is there?

Get a taxonomy report from PSM report and Unipept pept2lca table

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVITGEK</td>
<td>Archaea</td>
<td>Nitrosopumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

SQLite Relational Data base

**SQL** query joins PSM and LCA to report number of PSMs and Peptides per genus

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDV...</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAAAVGTVPE...</td>
<td>VAAAVGTVPEQEWLK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>genus</th>
<th>PSMs</th>
<th>DISTINCT PEPTIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planktomarina</td>
<td>161</td>
<td>20</td>
</tr>
<tr>
<td>Nitrosopumilus</td>
<td>122</td>
<td>27</td>
</tr>
</tbody>
</table>
Who is there? What are they doing?

How do we get taxonomy and function of a microbiome from a list of peptides?

<table>
<thead>
<tr>
<th>PSM#</th>
<th>Proteins</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EGGEDMFVHKSDVDGFINEGDK</td>
<td>EGGEDMFVHK</td>
</tr>
<tr>
<td>2</td>
<td>GKRVAAVGTVEQEWLK, KVAAAVGTVEQEWLK, RVAAAVGT...</td>
<td>VAAAVGTVEQEWLK</td>
</tr>
</tbody>
</table>

Pept2lca
taxonomy lowest common ancestor for a peptide

Pept2prot
Uniprot entries for a peptide with GO terms

<table>
<thead>
<tr>
<th>peptide</th>
<th>superkingdom</th>
<th>genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADGHTMHFDVITGEK</td>
<td>Archaea</td>
<td>Nitrosopumilus</td>
</tr>
<tr>
<td>AALESFTGNVTSALK</td>
<td>Bacteria</td>
<td>Polaribacter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peptide</th>
<th>uniprot_id</th>
<th>go_references</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAEKLAAQAR</td>
<td>W5T6F9</td>
<td>GO:0016021</td>
</tr>
<tr>
<td>AAEKLAAQAR</td>
<td>A0A0Q6ZKK0</td>
<td>GO:0005524, GO:0016887, GO:0015833</td>
</tr>
</tbody>
</table>
**QUESTIONS**

- How can I match metaproteomic mass spectrometry data to peptide sequences derived from shotgun metagenomic data?

- How can I perform taxonomy analysis and visualize metaproteomics data?

- How can I perform functional analysis on this metaproteomics data?

[https://training.galaxyproject.org/training-material//topics/proteomics/tutorials/metaproteomics/tutorial.html](https://training.galaxyproject.org/training-material//topics/proteomics/tutorials/metaproteomics/tutorial.html)
Metaproteomics Workflow

**Database Generation**
- FASTQ
- Protein / Peptide FASTA
- Spectra

**Database Search & Strategies**
- Search Algorithm
- Peptides

**Functional Analysis**
- Known Function
- Proteins
  - Hypothetical Function
  - Unknown Function

**Taxonomy Analysis**
- Unique Peptides
- Shared Taxonomy
  - Unassigned Taxonomy
Metaproteomics Workflow

**DATABASE GENERATION**
- FASTQ
- Protein / Peptide FASTA
- Search Algorithm
- Spectra

**DATABASE SEARCH & STRATEGIES**
- Peptides

**QUANTITATIVE ANALYSIS**
- Spectral counts OR Intensity data

**FUNCTIONAL ANALYSIS**
- Known Function
- Hypothetical Function
- Unknown Function
- Shared Taxonomy
- Unassigned Taxonomy

**TAXONOMY ANALYSIS**
- Unique Peptides
metaQuantome

Function: Volcano plots
Fold-change: 33 hours versus 8 hours

Function: Heatmap

metaQuantome on Galaxy Training Network

- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-data-creation/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-data-creation/tutorial.html)
- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-function/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-function/tutorial.html)
- [https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-taxonomy/tutorial.html](https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/metaquantome-taxonomy/tutorial.html)
METATRANSCRIPTOMICS ON GALAXY TRAINING NETWORK (ASaiM-MT)

https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/metatranscriptomics/tutorial.html

ASaiM
Batut et al Gigascience. 2018 7(6)
doi: 10.1093/gigascience/giy057


Accessing tools and Workflows

**METAGENOMICS:**
Toolshed: [z.umn.edu/metagenomics_toolshed](https://z.umn.edu/metagenomics_toolshed)
Galaxy Training Network: [https://training.galaxyproject.org/training-material/topics/metagenomics/](https://training.galaxyproject.org/training-material/topics/metagenomics/)

**METATRANSCRIPTOMICS:**
Workflow: [http://z.umn.edu/MTWF2020](http://z.umn.edu/MTWF2020)

**METAPROTEOMICS:**
Workflow: [z.umn.edu/MPWF2020](http://z.umn.edu/MPWF2020)
Galaxy Training Network: [http://z.umn.edu/gtn-metaproteomics](http://z.umn.edu/gtn-metaproteomics)

Also available on: [https://proteomics.usegalaxy.eu/](https://proteomics.usegalaxy.eu/) and Metaproteomics Gateway: [z.umn.edu/metaproteomicsgateway](http://z.umn.edu/metaproteomicsgateway)

galaxyp.org/contact