A “Divide and Conquer” approach to address peptide-spectrum matching challenges of large sequence databases in next-generation proteomic applications

Praveen Kumar
Challenge in PG and MP

- Database size
- Composition of the database

*PG - Proteogenomics
*MP - Metaproteomics
Challenge in PG and MP

- Database size
- Composition of the database

*PG - Proteogenomics
*MP - Metaproteomics
Challenge in PG and MP

- Database size
- Composition of the database

*PG - Proteogenomics
*MP - Metaproteomics

https://comic.browserling.com/extra-needle-haystack-raw.png
Proteomics Database Size

Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data

Dhirendra Kumar, Amit Kumar Yadav, and Debasis Dash

3 What Makes the Best Proteomics Search Database

Database constitutes arguably the most influencing factor in the protein discovery from tandem mass spectrometry data. Despite its importance, the choice of the database is to be evaluated before deciding on the search database for a given proteomic data. While organism-specific reference proteomes should be preferred choice over global protein databases, differences among reference proteomes from different sources present a complicated context-specific use case for MS/MS data analysis. A small exercise with only a fraction of overall data to evaluate which reference proteome maximizes the expected results, should facilitate an informed decision on the optimal search database. For targeted studies, like novel gene or isoform discovery, custom database should be designed. However, while designing such dedicated databases, a balance between comprehensiveness and compactness should be maintained.
Nature of Databases

- Proteogenomics (PG)
  - Generated from genomic and/or transcriptomic data
    - CustomProDB (Wang X. et al., 2013)
    - translation
    - splice junction

- Metaproteomics (MP)
  - Generated from genomic and/or transcriptomic data
    - Sixgill (May D.H. et al., 2016)
    - MOCAT (Kultima J.R. et al., 2016)
    - Graph2Pro (Tang H. et al., 2016)

- Proteome sequences from repositories
Common Approaches in Database Searching

- Traditional database search (Large DB)
  - Low sensitivity
- Two-step database searching method
  - May bias the database composition
  - Requires validation
  - Jagtap et al., 2013
    - DOI: 10.1002/pmic.201200352
- Other Methods:
  - COMPIL 2.0 (Chatterjee S. et al., 2016)
  - ProteoStorm (Beyer D. et al., 2018)
  - Graph2Pro-var (Li S. et al., 2019)
Common Approaches in Database Searching

- Traditional database search (Large DB)
  - Low sensitivity
- Two-step database searching method
  - May bias the database composition
  - Requires validation
  - Jagtap et al., 2013
    - DOI: 10.1002/pmic.201200352
- Other Methods:
  - COMPIL 2.0 (Chatterjee S. et al., 2016)
  - ProteoStorm (Beyter D. et al., 2018)
  - Graph2Pro-var (Li S. et al., 2019)
Common Approaches in Database Searching

• Traditional database search (Large DB)
  • Low sensitivity

• Two-step database searching method
  • May bias the database composition
  • Requires validation
  • Jagtap et al., 2013
    – DOI: 10.1002/pmic.201200352

• Other Methods:
  • COMPIL 2.0 (Chatterjee S. et al., 2016)
  • ProteoStorm (Beyter D. et al., 2018)
  • Graph2Pro-var (Li S. et al., 2019)
Common Approaches in Database Searching

- Traditional database search (Large DB)
  - Low sensitivity
- Two-step database searching method
  - May bias the database composition
  - Requires validation
  - Jagtap et al., 2013
    - DOI: 10.1002/pmic.201200352
- Other Methods:
  - COMPIL 2.0 (Chatterjee S. et al., 2016)
  - ProteoStorm (Beyter D. et al., 2018)
  - Graph2Pro-var (Li S. et al., 2019)
Common Approaches in Database Searching

- **Traditional database search (Large DB)**
  - Low sensitivity
- **Two-step database searching method**
  - May bias the database composition
  - Requires validation
  - Jagtap et al., 2013
    - DOI: 10.1002/pmic.201200352
- **Other Methods:**
  - COMPIL 2.0 (Chatterjee S. et al., 2016)
  - ProteoStorm (Beyter D. et al., 2018)
  - Graph2Pro-var (Li S. et al., 2019)
Divide and Conquer Approach

• Sectioning method (modified from two-step)
  • Improve on the database composition
  • Increase true hits
  • Controls false positives

• Provide a robust workflow for large database searches (Galaxy)
  – https://z.umn.edu/sectioningWF

• Flexible to modify parameters as per use case
Sectioning Flowchart

Large database
Sectioning Flowchart

Large database

Each Sectioned Database

MS/MS Data

SearchGUI

PeptideShaker

Target Seq.

Yes

No

Discard
Sectioning Flowchart

- Large database
- Each Sectioned Database
- MS/MS Data
- SearchGUI
- PeptideShaker
- Target Seq.
- Qualifying Score Cutoff (Optional)
- Yes: Target Sequences (n)
- No: Discard
- Select Random sequences equal to number of selected Target (n)
- Enriched Database (2n)
- Sequences from each sectioned database search
- SearchGUI
- PeptideShaker
- Proteins and Peptides to report
- PSMs, Peptides and Proteins to report
Sectioning Flowchart

1. Large database
2. Each Sectioned Database
3. SearchGUI
4. PeptideShaker
5. Target Seq.
6. MS/MS Data
7. Qualifying Score Cutoff (Optional)
8. Target Sequences (n)
9. Enriched Database (2n)
10. PSMs, Peptides, and Proteins to report

- Yes, proceed to Target Sequences (n)
- No, discard
- Select Random sequences equal to number of selected Target (n)
- Sequences from each sectioned database search
- Yes, proceed to SearchGUI
- No, discard

www.galaxyp.org
www.galaxyproject.org
Galaxy Workflow

https://z.umn.edu/sectioningWF
Systematic Evaluation of the Sectioning Method
Evaluation Datasets

**Pyrococcus furiosus (Pfu) Data**
- *Pfu*, a hyperthermophile
  - evolutionarily distinct from human
  - shares only 7 tryptic peptides with human
  - Vaudel, M et al., 2012 (10.1021/pr300055q)
  - Wong, CC et al., 2013 (10.1021/pr300840j)

- Mimic large database
  - *Pfu* protein sequences (2045 proteins)
  - Human protein sequences (93440 proteins)

**SIHUMI (simplified human intestinal microbiota) Data**
- 3rd International Metaproteomics Symposium (Dec. 2018)
  - [https://z.umn.edu/ims_data](https://z.umn.edu/ims_data)

- 8 organisms
  - *Anaerostipes caccae*
  - *Bacteroides thetaiotaomicron*
  - *Bifidobacterium longum*
  - *Blautia producta*
  - *Clostridium butyricum*
  - *Clostridium ramosum*
  - *Escherichia coli*
  - *Lactobacillus plantarum*

- Mimic large database
  - SIHUMI sequences (29,520 proteins)
  - Archaea sequences (1.47 million proteins)
Enriched Database:

- Standard + Entrapment proteins
- Traditional (Large Single Database)
- Traditional Two-Step

SearchGUI/PeptideShaker (SG/PS)

- Database: Standard + Entrapment proteins
- x Sections
- SG/PS
- PSMs

- Database: Standard + Entrapment proteins
- y Sections
- SG/PS
- PSMs

- Database: Standard + Entrapment proteins
- z Sections
- SG/PS
- PSMs

Database:

- Standard proteins
- Traditional (Large Single Database)

Compare the effect of sectioning method with traditional methods

MS/MS Spectra

- Enriched Database
- Enriched Database
- Enriched Database
- Enriched Database

PSMs

1% FDR
Database:
Standard proteins

Traditional (Large Single Database)

PSMs

1% FDR

MS/MS Spectra

SearchGUI/PeptideShaker (SG/PS)

Database:
Standard + Entrapment proteins
(Traditional Two-Step)

x Sections

SG/PS

Database:
Standard + Entrapment proteins

y Sections

SG/PS

Database:
Standard + Entrapment proteins

z Sections

SG/PS

Enriched Database

Enriched Database

Enriched Database

Enriched Database

Compare the effect of sectioning method with traditional methods
Enriched Database:

- Standard + Entrapment proteins

Traditional (Large Single Database)

- Standard + Entrapment proteins

Two-Step method

Database:

- Standard proteins
- Standard + Entrapment proteins

SearchGUI/PeptideShaker (SG/PS)

Compare the effect of sectioning method with traditional methods.
Enriched Database

Database:
Standard + Entrapment proteins

Traditional (Large Single Database)

Standard + Entrapment proteins

(Traditional Two-Step)

SG/PS

SearchGUI/PeptideShaker (SG/PS)

Two-Step method

x Sections

y Sections

z Sections

Compare the effect of sectioning method with traditional methods

MS/MS Spectra

Database:
Standard proteins

PSMs 1% FDR

MS/MS Spectra

Enriched Database

Database:
Standard proteins PSMs

Enriched Database

Database:
Standard proteins

Enriched Database

Database:
Standard proteins
Compare the effect of sectioning method with traditional methods
Results from the *Pfu* data

**Pyrococcus furiosus (Pfu) Data**

- *Pfu*, a hyperthermophile
  - evolutionarily distinct from human
  - shares only 7 tryptic peptides with human
  - Vaudel, M et al. (2012) (10.1021/pr300055q)
  - Wong, CC et al. (2013) (10.1021/pr300840j)

- **Pfu PSMs:** True
- **Human PSMs:** False

- No. of Sections: 2, 5, and 10

**SIHUMI (simplified human intestinal microbiota) Data**

  - [https://z.umn.edu/ims_data](https://z.umn.edu/ims_data)
- 8 organisms
  - Anaerostipes caccae
  - Bacteroides thetaiotaomicron
  - Bifidobacterium longum
  - Blautia producta
  - Clostridium butyricum
  - Clostridium ramosum
  - Escherichia coli
  - Lactobacillus plantarum

- Mimic large database
  - SIHUMI sequences (29,520 proteins)
  - Archaea sequences (1.47 million proteins)
Results from the \textit{Pfu} data

\begin{figure}
\centering
\includegraphics[width=\textwidth]{results.png}
\caption{Comparison of number of \textit{Pfu} PSMS and number of Human PSMS across different methods and section counts.}
\end{figure}
Results from the SIHUMI data

**Pyrococcus furiosus (Pfu) Data**
- *Pfu*, a hyperthermophile
  - evolutionarily distinct from human
  - shares only 7 tryptic peptides with human
  - Vaudel, M et al. (2012) (10.1021/pr300055q)
  - Wong, CC et al. (2013) (10.1021/pr300840j)
- Mimic large database
  - *Pfu* protein sequences (2045 proteins)
  - Human protein sequences (93440 proteins)

**SIHUMI (simplified human intestinal microbiota) Data**
  - https://z.umn.edu/ims_data
- 8 organisms
  - *Anaerostipes caccae*
  - *Bacteroides thetaiotaomicron*
  - *Bifidobacterium longum*
  - *Blautia producta*
  - *Clostridium butyricum*
  - *Clostridium ramosum*
  - *Escherichia coli*
  - *Lactobacillus plantarum*
- SIHUMI PSMs: True
- Archaea PSMs: False
- No. of Sections: 5, 10, 20, and 30
Archaea PSMs and Confidence Score

Percent = \frac{Archaea PSMs}{(SIHUMI PSMs + Archaea PSMs)} \times 100

Confidence Scores

Intervals of 5

Intervals of 1
Database Size After Reduction

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (SIHUMI)</td>
<td>29,635</td>
</tr>
<tr>
<td>Traditional Large (SIHUMI + Archaea)</td>
<td>1,472,421</td>
</tr>
<tr>
<td>Traditional Two-Step</td>
<td>17,179</td>
</tr>
<tr>
<td>5-sections</td>
<td>238,191</td>
</tr>
<tr>
<td>10-sections</td>
<td>396,255</td>
</tr>
<tr>
<td>20-sections</td>
<td>578,785</td>
</tr>
<tr>
<td>30-sections</td>
<td>723,243</td>
</tr>
</tbody>
</table>
Glacial Meltwater Microbiome Data

- Proteins extracted from the microbiome of different layers of melting water – *(Nunn Lab, University of Washington)*

- Database **generated using metagenomics data** (using Sixgill and MOCAT)

- Database size: **5,117,895** sequences (Traditional large DB search)

- Reduced database size (using 5-Sections): **204,649** sequences

Unipept Analysis

Name this search: optional
- Equal I and L
- Filter duplicate peptides
- Advanced missed cleavage handling

Enter a peptide list
Paste a list of peptides from your metaproteomics experiment in the search box. Each line should contain a single peptide.

Click search
Review the search settings for peptide deduplication, advanced missed cleavage handling and equating isoleucine and leucine before clicking search.

Gain insight
For each of the peptides, the lowest common ancestor (LCA) will be calculated, aggregated and visualised on the result page, giving you insight in the taxonomic diversity of your sample.

Protein EC Result

Assigned Enzyme Commission Number (EC)

Traditional

Sectioning

10

346

205

Top 15 Protein Enzyme Groups (EC)

- H+ transporting two-sector ATPase
- RNA-dependent RNA polymerase
- Protein-synthesizing GTPase
- Glutamine synthetase
- Adenosine triphosphatase
- Methane monooxygenase (soluble)
- Photosystem I
- Hydrolases-Acting on acid anhydrides
- 2-alkenal reductase (NAD(P)(+))
- Beta-ketoacyl-[acyl-carrier-protein] synthase

Number of PSMs
Gene Ontology

Top 15 Biological Processes (PSM assignments)

- Protein refolding
- ATP synthesis coupled proton transport
- Translation
- Transcription, DNA-templated
- Photosynthesis
- Reductive pentose-phosphate cycle
- Glycolytic process
- Gluconeogenesis
- Protein transport
- Glutamine biosynthesis process
- Protein import
- Protein dephosphorylation
- RNA binding
- Protein kinase activity
- DNA-directed RNA polymerase activity
- DNA-directed RNA polymerase II promoter activity

Top 15 Molecular Functions (PSM assignments)

- ATP binding
- Unfolded protein binding
- Protein-transporting ATP synthase activity
- DNA binding
- Nucleoside diphosphate kinase activity
- Mitochondrion
- Translation elongation factor activity
- Metal ion binding
- NADH dehydrogenase activity
- NADH dehydrogenase (ubiquinone) Fe-S oxidoreductase activity
- Mitochondrion
- RNA binding
- GTPase activity
- DNA-directed RNA polymerase II process
- DNA-directed RNA polymerase II promoter activity
- Translation elongation factor activity
- DNA polymerase activity

Number of PSMs
Conclusions

- Identifies more PSMs than the traditional large database search method
- Reduces the identification of false positives compared to the traditional two-step method
- More number of sections yield high confidence PSMs
- Improves the composition of the sequences in the enriched database
- **CAUTION:**
  - Identification: PSM score cutoff recommended
  - Quantification: Optimal database size with stricter PSM score cutoff
- Flexible Galaxy workflow ([https://z.umn.edu/sectioningWF](https://z.umn.edu/sectioningWF))
Available workflow and Future Directions

- Flexible Galaxy workflow ([https://z.umn.edu/sectioningWF](https://z.umn.edu/sectioningWF))
Available workflow and Future Directions

• Flexible Galaxy workflow (https://z.umn.edu/sectioningWF)

• Apply to more metaproteomic datasets

• Apply and evaluate using proteogenomics datasets

• Coming soon: Docker container image
Acknowledgements

- **Advisor:** Dr. Timothy Griffin
- Dr. Pratik Jagtap
- **Galaxy-P Team members:**
  - James Johnson
  - Thomas McGowan
  - Subina Mehta
  - Caleb Easterly
  - Ray Sajulga
  - Andrew Rajczewski
  - Dr. Shane Hubler
  - Marie Crane
  - Emma Leith
- **BICB, University of Minnesota**
- Dr. Joel Rudney (University of Minnesota)
- Dr. Brook Nunn (University of Washington)
- Dr. Robert Hettich (Oak Ridge National Laboratory)
- Dr. Nico Jehmlich (The Helmholtz-Centre for Environmental Research)
- **Galaxy Community (Worldwide)**
- **Computational Infrastructure**
  - Minnesota Supercomputing Institute
  - European Galaxy [https://usegalaxy.eu](https://usegalaxy.eu)
- **Funding**
  - Norwegian Centennial Chair (NOCC)
  - National Science Foundation (NSF)
  - National Institutes of Health (NIH)
  - ASMS Graduate Travel Award
Thank you!