Spectral Counting Approaches and PEAKS

INBRE Proteomics Workshop, April 5, 2017
Boris Zybailov
Department of Biochemistry and Molecular Biology
University of Arkansas for Medical Sciences

1. Introduction to PEAKS software
   1.1 Data Pre-processing and De novo sequencing
   1.2 PEAKS database search
   1.3 PEAKS Spider
   1.4 Peptide mapping and PTM analysis in PEAKS
   1.5 InChorus, integration with other proteomics workflows
   1.6 Data export options. Export to Scaffold
   1.7 PEAKS Q module: TMT-labeling workflow and label free quantification

2. Spectral Counting
   2.1 Applicability
   2.2 Spectral counting vs. IBAC and integrated XIC methods
   2.3 What statistical tests are appropriate for the Spectral Count data
   2.4 Inter-sample and intra-sample variation. What can be inferred from one biological replicate?

3. Metaproteomics Analysis workflow using PEAKS, Spectral Counting, and TMT-labeling
PEAKS Studio is The Desktop Proteomics Software Suite, Available from Bioinformatics Solutions

The key feature: *de novo* interpretation of MS/MS spectra

LC-MS- MS/MS → data refinement → De novo tags → Database Search

To get started http://www.bioinfor.com/walkthrough-tutorials/
• Multiple vendor-specific formats are supported for input (we will focus on Thermo.RAW)
• MGF input is supported
• mzXML is supported
PEAKS: Data Refinement

• Low quality MS/MS data: \( q > 0.65 \) is recommended

• Typical Orbitrap-Ion-trap or Orbitrap-Orbitrap experiment - no filtering of scans is needed

• Example: 60,000 to 240,000 resolving power in Orbitrap, HCD fragmentation, MS/MS in ion-trap, \( \pm 0.5 \) Da fragment tolerance

• Data Refinement is done automatically and is not optional in the latest version of PEAKS Studio (v 8.0)

• Recalculation of multiply-charged fragments improves de novo quality

• In the latest version, X!Calibur is not required for Thermo files
PEAKS: Data Input and Refinement
PEAKS: De Novo Algorithm

- using platform-specific fragmentation model find the best sequence tags
- masses of the tags are the same as the precursor, given tolerance
- Use a, b, c, x, y, z, b-H2O, b-NH3, y-H2O, y-NH3, terminal, internal cleavage, and immonium ions

Original PEAKS paper:


14 years later: everything works faster, New features are available, essential de novo algorithm is the same
Most of the details are proprietary,

Most important steps can be inferred from the original and subsequent publications

**PEAKS: Optimal Configuration**

32 computing threads
Desktop, Windows 7

Processor: Intel(R) Xeon(R) CPU E5-2657 v4 @ 2.30GHz  2.30 GHz
Installed memory (RAM): 128 GB
System type: 64-bit Operating System

**PEAKS: De Novo Speed**

80 – 150 MS/MS spectra per second,
~1 hr for 300,000 MS/MS spectra – a typical gel lane, 24 slices

The second step, candidate computation, is the critical step in which the 10000 best sequences of all possible combinations of amino acids for a given precursor ion mass are computed. For this computation, the a, b, c, x, y and b/y –17/18 ions are considered. The basic assumption of our model is that the greater number of high abundance peaks that are matched by those ions of a sequence, the more likely the predicted sequence is the correct sequence. For each mass value \( m \), this new algorithm first computes the reward/penalty that a y (or b) ion has mass \( m \). If there is a peak close to \( m \), the reward is equal to the logarithmic abundance of the peak multiplied by a factor reflecting the mass error between \( m \) and the mass value of the peak, and multiplied by a factor reflecting the co-existence of the x, y-H\(_2\)O, y-NH\(_3\) (or a, c, b-H\(_2\)O, b-NH\(_3\)) ions. If there is no peak close to \( m \), the reward is a negative constant value. The problem is then reduced to finding a sequence such that its y and b ions maximize the total rewards at their mass values.

The initial mathematical formula used to compute the
Trypsin specificity for de novo tags implies tryptic C-termini.
For Trypsin and/or Endo-LysC digests setting to no enzyme will produce lower number of tags.

Low accuracy (fragments acquired in ion trap, 0.5 Da tolerance) does not slow the De Novo speed.

Parameters affecting the de novo speed.
In this example,
Two population of high quality and low-quality De novo tags are evident

Average Local Confidence Score (ALC) ranges from (15-100%)

Local confidence score changes from 0 to 100%

Table 1. Statistics of data and result.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of MS scans</td>
<td>0</td>
</tr>
<tr>
<td># of MS/MS scans</td>
<td>324179</td>
</tr>
<tr>
<td>Peptides after filtration</td>
<td>183791</td>
</tr>
</tbody>
</table>

Table 2. Result filtration parameters.

<table>
<thead>
<tr>
<th>ALC score</th>
<th>≥50%</th>
</tr>
</thead>
</table>
**Local Confidence** is the confidence that a particular amino acid is present in the de novo peptide at a particular position. It is presented as a percentage.

**Total Local Confidence (TLC)** is the sum of the local confidence scores (0 to 100%) from each amino acid in the peptide sequence.

**Average Local Confidence (ALC)** is the average of the TLC. It is TLC divided by the number of amino acids in the peptide sequence.
PEAKS: Average Local Confidence Score, %
PEAKS: Average Local Confidence Score, %
PEAKS: Average Local Confidence Score, %
PEAKS: Database Search

- No support for the new Change in NCBI database just yet,
- Only old gi numbers are supported –
- Relevant for taxonomy;
- Parsing rules can be customized
In the current version

“No Cleavage”
Does not exactly work in PEAKS
PEAKS: Database Search

Peaks’ “Decoy-Fusion” method
Allows for correct estimation of false positives rate
Through a multi-step database search
PEAKS: Decoy-Fusion method – works well for multi-step Database Search

Example for metaproteomics of the gut analysis:

- Search full uniprot/tremble database, restricted by host taxonomy
- Transfer unmatched de novo tags to a new step, search against food protein database
- Transfer unmatched de novo tags from the previous step and search against all bacteria
- Prepare custom fasta database of identified proteins across different steps
- Re-search the full dataset against the custom database
PEAKS: Result Summary

1. Notes

2. Result Statistics

Figure 1. False discovery rate (FDR) curve. X axis is the number of peptide-spectrum matches (PSM) being tested. Y axis is the corresponding FDR.

Figure 2. PSM score distribution. (a) Distribution of PEAKS peptide score; (b) Scatterplot of PEAKS peptide score versus precursor mass error.
PEAKS: Result Summary

Table 1. Statistics of data.
- # of MS scans: 159724
- # of MS/MS scans: 311007

Table 2. Result filtration parameters.
- Peptide -10gP: ≥20.3
- Protein -10gP: ≥20
- Proteins unique peptides: ≥2
- De novo ALC Score: ≥70%

Table 3. Statistics of filtered result.
- Peptide-Spectrum Matches: 142630
- Peptide sequences: 53509
- Protein groups: 3173
- Protein: 11041
- Proteins (#Unique Peptides): 8851 (≥2); 2180 (=2); 0 (=1);

Table 4. PTM profile.

<table>
<thead>
<tr>
<th>Name</th>
<th>Δ Mass</th>
<th>Position</th>
<th>#PSM</th>
<th>-10gP</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbamidomethyl</td>
<td>57.02</td>
<td>C</td>
<td>23903</td>
<td>128.45</td>
<td>2.9</td>
</tr>
<tr>
<td>Oxidation</td>
<td>15.93</td>
<td>M</td>
<td>16875</td>
<td>128.45</td>
<td>2.1</td>
</tr>
<tr>
<td>Deamidation</td>
<td>.98</td>
<td>NQ</td>
<td>11119</td>
<td>119.92</td>
<td>7.1</td>
</tr>
<tr>
<td>Acetylation</td>
<td>42.01</td>
<td>N-term</td>
<td>4562</td>
<td>117.54</td>
<td>3.2</td>
</tr>
<tr>
<td>Carbamidomethyl</td>
<td>57.02</td>
<td>DEHK,N-term</td>
<td>1593</td>
<td>129.34</td>
<td>2.9</td>
</tr>
<tr>
<td>HexNAc</td>
<td>203.08</td>
<td>N</td>
<td>1242</td>
<td>84.46</td>
<td>2.1</td>
</tr>
<tr>
<td>Methyl ester</td>
<td>14.02</td>
<td>DE,C-term</td>
<td>1207</td>
<td>96.33</td>
<td>5.1</td>
</tr>
<tr>
<td>Dihydroxy</td>
<td>31.93</td>
<td>CFKPRWY</td>
<td>586</td>
<td>108.81</td>
<td>5.1</td>
</tr>
<tr>
<td>Oxidation</td>
<td>15.99</td>
<td>HW</td>
<td>878</td>
<td>113.25</td>
<td>2.1</td>
</tr>
<tr>
<td>Pyro-glu from Q</td>
<td>-17.03</td>
<td>N-term</td>
<td>855</td>
<td>89.23</td>
<td>8.2</td>
</tr>
<tr>
<td>Ala&gt;Pro</td>
<td>26.02</td>
<td>P</td>
<td>469</td>
<td>85.56</td>
<td>3.1</td>
</tr>
<tr>
<td>Methylation</td>
<td>14.02</td>
<td>HKST,N-term</td>
<td>464</td>
<td>105.09</td>
<td>1.1</td>
</tr>
<tr>
<td>Sulfates</td>
<td>91.98</td>
<td>NT,C-term</td>
<td>345</td>
<td>87.77</td>
<td>7.1</td>
</tr>
</tbody>
</table>
### Table 3. Statistics of filtered result.

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide-Spectrum Matches</td>
<td>142630</td>
</tr>
<tr>
<td>Peptide sequences</td>
<td>53509</td>
</tr>
<tr>
<td>Protein groups</td>
<td>3173</td>
</tr>
<tr>
<td>Proteins</td>
<td>11041</td>
</tr>
<tr>
<td>Proteins (#Unique Peptides)</td>
<td>8861 (&gt;2); 2180 (=2); 0 (=1);</td>
</tr>
<tr>
<td>FDR (Peptide-Spectrum Matches)</td>
<td>1.0%</td>
</tr>
<tr>
<td>FDR (Peptide Sequences)</td>
<td>1.8%</td>
</tr>
<tr>
<td>FDR (Protein)</td>
<td>0.5%</td>
</tr>
<tr>
<td>De Novo Only Spectra</td>
<td>40900</td>
</tr>
</tbody>
</table>
SPIDER – A homology search tool that is designed to match de novo tags directly to the database. This allows to identify peptides even when working with an unsequenced organism or highly variable proteins.

Figure 1: Illustrating how SPIDER reconstructs the correct sequence from a de novo sequence and a homologous peptide.
a very useful application of SPIDER is to use it iteratively to sequence a complete protein (e.g. antibody sequencing). This is achieved by:

• Using PEAKS’ standard workflow (de novo + PEAKS DB + PEAKS PTM + SPIDER) to search in a homologous database. This will identify a homologous protein.

• Then in the coverage pane, select tools “copy mutated protein sequence”. This will copy the mutated protein sequence (after applying the confident mutations) to Windows’ clipboard.

• Invoke another standard search by paste the copied sequence as the protein database.

• Repeat the above procedure multiple times to gradually improve the sequence quality.
Peptide Mapping

Annotated Chromatogram of Protein sp\,|\,P12821\,|\,ACE_HUMAN

Fraction Choices: F 7, F 8, F 9

Minimal Ion Intensity: 1E4

Display Mode: PEAKS DB, Annotation

RT: 10, 20, 30, 40, 50, 60, 70, 80

Intensity: 3.75E7, 1.87E7
Peptide Comparison of Protein sp | P12821 | ACE_HUMAN
PTM profiling

Protein: Q96QL8|Q96QL8_HUMAN
Fraction: F2
Modification: All the PTMs

All PTMs on Q96QL8|Q96QL8_HUMAN

F2
(Unmodified NT22d) = 93%

F7

Modified
Unmodified

Position

PEAKS: InChorus – support for Multiple Search Engines

- Raw MS/MS Spectra
- de Novo
  - Database Search
    - Mascot
    - OMSSA
    - Sequest
  - PTM Finder
  - Seq.Tag Homology Search
- inChorus (Summary + Consensus + Individual Reports)
PEAKS: InChorus – support for Multiple Search Engines
Figure 1: The false discovery rate (FDR) curve. X-axis is the number of PSMs (peptide-spectrum matches) being kept and the y-axis is the FDR.

Figure 2: Venn diagram for confident PSMs (up to three selected search engines).

<table>
<thead>
<tr>
<th>Engine</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEAKS</td>
<td>1657(20)</td>
<td>1.2%</td>
</tr>
<tr>
<td>Mascot</td>
<td>90(1)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Mascot</td>
<td>133(1)</td>
<td>8.3%</td>
</tr>
<tr>
<td>Mascot</td>
<td>102(4)</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Mascot</td>
<td>199(0)</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Mascot</td>
<td>72(0)</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Mascot</td>
<td>232(6)</td>
<td>3.9%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>1024(0)</td>
<td>8.3%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>232(6)</td>
<td>3.9%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>1657(20)</td>
<td>1.2%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>90(1)</td>
<td>0.1%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>133(1)</td>
<td>8.3%</td>
</tr>
<tr>
<td>Target(left)</td>
<td>1024(0)</td>
<td>8.3%</td>
</tr>
</tbody>
</table>
PEAKS Q

No Support for Spectral Counts,
But it is possible to export results into Scaffold

Advanced support for stable-isotope labeleing, including
TMT10-plex MultiNotch

Moderate support for XIC integration and comparison
Metaproteomics Pipeline

Sample preparation
- Fecal microbiome sample
- Collect an aliquot (~1 g). Split in 4 equal parts at the time of collection and freeze at -80 °C

Metaproteomics Analysis
- Protein extracts
  1) SDS-Page in gel digestion
  2) On-filter digestion

LC-MS
- Orbitrap-Tribrid-Fusion
  1) Identification method
  2) Quantification (TMT tags)

Metagenomics
- Illumina HiSeq 2000/2500
  a) Assembly using reference genomes
  b) De novo assembly

Sequence databases
- Sample-derived DNA sequence database
- NCBI_nr, taxonomy filters
- host+microbiome protein database

Metagenomics
- a) Assembly using reference genomes
- b) De novo assembly

Sequence databases
- Sample-derived DNA sequence database
- NCBI_nr, taxonomy filters
- host+microbiome protein database

Data cross-validation
- De Novo Only peptides
- peptides matched to proteins

Results:
1) List of organisms (species-level resolution) and their relative abundance
2) List of proteins and their abundance ranks
3) Matched peptides and de novo only peptides relative concentrations
Multi Step Database Search Strategy

Microbiome Sample

Label Free Method
Gel-LC20
Fusion

.RAW

PEAKS 8.0
DeNovo
5 ppm
0.5 Da

DeNovo peptides

ALC (%) ≥ 80

DeNovo Peptides, 80 Filtered list

UniPept – web
advanced miss-cleavage
Equate L and I
Do not filter duplicates

Taxonomy/diversity
Exact matches in UniProt

PEAKS DB
NCBI_nr (host taxonomy)
Strict filtering

Unmatched DeNovo tags

PEAKS DB
NCBI_nr (bacteria)
Relaxed filtering

Prepare Custom Fasta Database
From all protein matched
The 410 quantified proteins separate the two phenotypes (CKD vs CKDRS) very well.

Outlier test: CKD9 and CKDRS21 – currently considering to re-run just these two rats.

T-test, 2 tails, unpaired, Equal variance

Distance function: Euclidean or correlation – Euclidean is shown on this plot

Clustering – complete linkage

Bioconductor package: “Heatplus”

107 proteins With adjusted p<0.05
107 most significant proteins (p.adjusted<0.05)

Clustering: centroid
Distance: correlation

Mostly rat proteins
Mostly bacterial proteins

Next slide
Shows the protein list
Conclusions

PEAKS outperforms most of the existing search algorithms in terms of speed, sensitivity and accuracy.

However, protein grouping and spectral counting support is inferior to Scaffold.

PEAKS can be used for preparing MGF files and De Novo tags for downstream analyses.