MS-GF+ and MSPathFinder for Identifying Peptides/Proteins

Sangtae Kim
Pacific Northwest National Laboratory
Mass Spec Data != Collection of MS/MS spectra

An experiment involves multiple LC-MS files.

People want to quantify proteins, not identify peptides.
MS-GF+

What happened after I left?

- mzML and mzIdentML support (Proteomics Standard Initiative standard formats)
- New scoring parameters for iTRAQ, TMT, Q-Exactive HCD, MHC peptides
- Lots of bug fixes / updates to meet users’ needs
MS-GF+ is being recognized by the community

I’ll show results from independent studies
MS-GF+ is worth a look if you haven't tried this free search engine yet bix-lab.ucsd.edu/pages/viewpage...
Pipeline/Post-processing Tools supporting MS-GF+

- Trans-Proteomic Pipeline, ISB
- Galaxy-P, U of Minnesota
- ProteoSuite, U of Liverpool
- ProteoSAFe/MassIVE, UCSD
- MSblender
- searchgui
- peptide-shaker
- Scaffold
- OPENMS
- IDPicker 3
- Bioconductor
- PepArML Meta-Search Engine
- Journal of Proteome Research
- Fast and Accurate Database Searches with MS-GF+Percolator
Database Search Engines @ PNNL

% Analysis Jobs

- Sequest
- MS-GF+
- X!Tandem

2011: 99% Sequest

2012: 9% X!Tandem

2013: 0% X!Tandem

Last 12 weeks: 0% X!Tandem
First, a comparison of the total number of PSM identifications showed that MS-GF+ added 67% more PSMs than Mascot-Percolator (Supplemental Fig. S3A).
The email that I received on 1/14/2014

“We do get amazing results from MS-GF+ when compared to Mascot or X!Tandem, maybe a little bit too good”

*Philipp Lange, Overall lab, University of British Columbia*
1,389 spectra
17 proteins
Waters Q-TOF

2,000 spectra
Human tumor tissue
Thermo Q-Exactive

2,000 spectra
Human tumor tissue
AB Sciex Triple-TOF 5600

5,806 spectra
Human blood plasma
Thermo LCQ

---

Risk et al., J. Proteome Res. 2013
"In essence, our approach is an application of the so-called “generating function” framework proposed by Kim et al. (4) to a well-known and easy-to-understand score function, the SEQUEST XCorr.

... In this study, the best overall statistical performances were clearly produced by the dynamic programming based score functions, MS-GF+ and our XCorr p-values."
I don’t know how you avoid the decoys so well but the difference is stunning… (Paul Rudnick)
Re-training for TMT Dramatically Increased #Identifications

#PSMs (1% FDR)

Fraction

MS-GF+ (Regular)
MS-GF+ (TMT)
MaxQuant

Fraction
“Finally, we have been training MSGF+ with our ASP-N data, its superb! We almost doubled our ID’s with the new training parameters, GREAT!!”
Present Focus
Top-Down and Bottom-Up Proteomics

Top-Down: intact proteins

Bottom-up: digested peptides
Data Independent Acquisition (DIA)

20 20 m/ξ-wide windows = 400 m/z

~2 seconds

~30 seconds

ASMS 2013 slides by Egertson et al.
MSPathFinder
**MASH Suite Visualization**

Ying Ge’s group developed MASH Suite, a user-friendly and versatile software interface for processing high-resolution mass spectrometry data. MASH Suite contains a wide range of features that allow users to easily navigate through data analysis, visualize complex high-resolution mass spectrometry data and manually validate automatically processed results. See paper here.

http://crb.wisc.edu/yinglab/software.html

**MS-Align+**

MS-Align+ is a software tool for top-down protein identification based on spectral alignment that enables searches for unexpected post-translational modifications. MS-Align+ is fast in identifying unexpected post-translational modifications. In addition, MS-Align+ reports statistical significance of top-down protein identifications.

http://bix.ucsd.edu/projects/msalign/

**MS-Deconv Deconvolution**

MS-Deconv is a software tool for top down spectral deconvolution. MS-Deconv uses a combinatorial algorithm. The algorithm first generates a large set of candidate isotopomer envelopes for a spectrum, then represents the spectrum as a graph, and finally selects its highest scoring subset of envelopes as a heaviest path in the graph. In contrast with other approaches, the algorithm scores sets of envelopes rather than individual envelopes.

http://bix.ucsd.edu/projects/msdeconv/

**ProSight PTM 2**

ProSight PTM 2.0 allows identification and characterization of both intact proteins and peptides. Our ProSight Warehouses are annotated with all known post-translational modifications (PTMs), alternative splicing events and single nucleotide polymorphisms (SNPs) using the technique of Shotgun Annotation developed in the Kelleher Research Group. ProSightPTM is the only proteomics software that allows the user to search their tandem MS data against proteome warehouses containing the known biological complexity present in UniProt.

https://prosightptm2.northwestern.edu/

**ProSightPC 3.0**

Thermo Scientific® ProSightPC, the first stand-alone software for analyzing top-down proteomics data, has been enhanced to add support for middle-down and bottom-up experiments, making it an all-around tool for identification and characterization of both intact proteins and peptides.

http://www.thermoscientific.com/ecom/ser

**YADA Deconvolution**

Description: YADA can deisotope and decharge high-resolution mass spectra from large peptide molecules, link the precursor monoisotopic peak information to the corresponding tandem mass spectrum, and account for different co-fragmenting ion species (multiplexed spectra). YADA also enables a pipeline consisting of ProLuCID and DTASelect for analyzing large-scale middle-down proteomics data.

http://fields.scripps.edu/yada/

http://www.topdownproteomics.org/software
To p-Down Software Tools

ProSightPC
“Annotated” proteoforms

MS-Align+

Proteoforms with blind PTMs (derived from a fasta file)

PathFinder
Proteoforms with known PTMs (derived from a fasta file)
Sequest/MS-GF+

Raw File

Fasta file

Modifications

Peptide-Spectrum Matches (PSMs)
“The first stand-alone software for analyzing top-down proteomics data…”

“…Search tandem MS data against proteome warehouses containing the known biological complexity present in UniProt…”
ProSightPC

Proteoform-Spectrum Matches (PrSMs)
Proteome Warehouse Approach

Good
Smaller search space → Less chance of spurious hits
More accurate PTM localizations

Bad
Available only for small organisms
Can never cover all possible proteoforms
Difficult to maintain
Discovery of novel information is limited
Protein Identification Using Top-Down*
Raw File → Fasta file → MS-Align+ → Proteoform-Spectrum Matches (PrSMs) with "blind" modifications (mass shifts)
MS-Align+

**Good**

Can discover novel PTMs / Mutations  
Efficiently handles the huge search space  
Reports E-values (similar to MS-GF+)

**Bad**

Results are not “clean”  
Excessive false positives
MS-Align+ Ids with false PTMs

The most creative ID so far:

Q.GSQVRLQVRVTGIPT(PVVKFYRD...GGLYTLSSLGNEFGSDS)\([-3796555.40]\)ATVNIHIRSI.-

Protein Mass: 3.8M Da

Discovered by Matt Monroe
Raw File - Fasta file - Modifications

PathFinder

Proteoform-Spectrum Matches (PrSMs)
Bottom-Up vs Top-Down

Why couldn’t I modify MS-GF+ to work for top-down?
Complexity of Spectra

Top-Down: More complex
(highly charged ions)

Bottom-Up: Less complex
Histone H3

#Proteoforms is too large!
(combinatorial explosion)
Histone H3

#Proteoforms ≈ $5 \cdot 10^{13}$

#Proteoforms is too large!
(combinatorial explosion)
Sequence Graphs
Precursor Product Pair (PPP)

PPP
A pair of (PrecursorIonComposition, ProductIonComposition)
Unit of scoring

#Unique PPPs << #Proteoforms*ProteinLength
Scoring a PPP (Informed)

(PrecursorIonComposition, ProductIonComposition)
Protein: MARTKQTARK \( (C_{48}H_{89}N_{19}O_{13}S) \)
Modification: Methyl K, \((CH_2)\), Oxidation M (O)
Max #Modifications: 2

Reversed Sequence

Possible modifications

- No modification
- 1 Methyl
- 1 Oxidation
- 2 Methyls
- 2 Oxidations
- 1 Methyl, 1 Oxidation

Compositions of possible proteoforms

- KmetRATQK
- KRATQKmet
\( (C_{31}H_{58}N_{12}O_{8}) \)

- “MARTKQTARK”, Methyl, Ox
\( (C_{49}H_{91}N_{19}O_{14}S) \)
Select “MARTKQTARK”, 1 Methyl, 1 Oxidation ($C_{49}H_{91}N_{19}O_{14}S$)

Each node represents a transition. Assign a score to each node. Find the best scoring path.
Sequence Graph

Naïve approach: $O(2^n)$ ($n$: sequence length)

With sequence graphs: $O(2^m)$ ($m$: #modifications)

For top-down proteomics,

$m << n$
Problem Solved?

I thought so, but…
Top-Down Proteomics is Trivial…?

#Proteins = 5.5K << #Peptides = 300K

55 times smaller!

Salmonella database
Not so trivial but still easy…?

#Proteins = 175K < #Peptides = 300K

Still ~2 times smaller!

Consider up to 30 N-term single residue cleavage (signal peptides)
Probably I was wrong!

There are many internal cleavages

#Proteins = 223M >> #Peptides = 300K

740 times larger
Even Worse

Protein

Peptide

Proteins are longer ➞ More fragments to score

Charges 1~4 for bottom-up, 2~30 for top-down
Speed matters!

I got new data. Can you run your tool?

No worries. It will take only 160 days. If you’re interested in finding PTMs, it takes a little more than 4 years. I wish your sample is not from human. If so, it won’t finish while you’re alive.
How to make it faster?
No obvious pattern!

QC_SheWIntact_2ug_3k_CID_4Apr14_Bane_PL011402.raw
MS-Align no modification search, FDR = 0
Are multiple cleavages common?
# Categorizing Proteoforms

<table>
<thead>
<tr>
<th>Protein</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No cleavage or N-term single residue cleavage</td>
<td>25%</td>
</tr>
<tr>
<td>Single internal cleavage (+ N-term single residue cleavage)</td>
<td>60%</td>
</tr>
<tr>
<td>Multiple cleavages</td>
<td>15%</td>
</tr>
</tbody>
</table>
IC Search Modes

Protein

- No cleavage or N-term single residue cleavage
- Single internal cleavage (+ N-term single residue cleavage)
- Multiple internal cleavages

Mode 2 (25%)
Mode 1 (85%)
Mode 0 (100%)
#Proteoforms

74X smaller  
27X smaller

<table>
<thead>
<tr>
<th>Mode 0</th>
<th>Mode 1</th>
<th>Mode 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>223M</td>
<td>3M</td>
<td>110K</td>
</tr>
</tbody>
</table>

These are numbers without considering PTMs
LC-MS Map

PrSMs are first discovered using a fast algorithm
And later re-scored using a more informed approach.
MS/MS #4824

Isolation window
[974.67, 977.67]

Previous MS1 #4816

Next MS1 #4825

Sum
Summed MS1 isolation window

Detect (multiple) features

Generate extracted ion chromatogram (XIC)

Register to “LC-MS Map”
Selecting Features

1. Pearson correlation between the Averagine should be >0.7

2. XIC correlation between the most abundant isotope and the next isotope should be >0.7

3. For charge c, the correlation between the XICs of the most abundant ions of c and c-1 or c+1 should be >0.7.
The image represents an LC-MS (Liquid Chromatography-Mass Spectrometry) map. It includes a selected feature with the most-abundant isotope indicated. The monoisotopic mass (M) is shown, and the LC range to be registered is noted. MS/MS with isolation windows includes (M-18)/charge.
Repeat this for all peptides in the database.
Deconvolution of MS/MS spectra

MS2 spectra are deconvoluted.

Find features where the correlation between the Average is above 0.7 and record monoisotopic masses.
Results
Running Time – No modification

MS-Align: ~1078 min
ProSightPC: ~1440 min

Mode 0
Decoy
Target
MS2Deconv
LcMsMap

Charles’s Salmonella dataset (SBEP_STM_001_02272012_Aragon)
Single threaded
#Protein Spectrum Matches (PrSMs)

No modification search, FDR 1%
Running Time – PTM Search

Charles’s Salmonella dataset (SBEP_STM_001_02272012_Aragon)
#PrSMs – PTM search

- IC (Mode 1)
- IC (Mode 2)
- MS-Align+

2 blind modifications

FDR 1%

Acetyl Prot N-term
Oxidation M
Dehydro C
Glutathione C
MSPathFinder for Bottom-Up
Probabilistic Scoring Model

Spectral Dictionaries
INTEGRATING DE NOVO PEPTIDE SEQUENCING WITH DATABASE SEARCH OF TANDEM MASS SPECTRA

Sangtae Kim‡, Nitin Gupta§, Nuno Bandeira‡, and Pavel A. Pevzner‡§

Christopher Wilkins implemented a probabilistic scoring model used by MS-GF+
Data sets

Mouse Heart Proteome

Q Exactive

DDA (2 m/z isolation)

DIA (5 m/z isolation): 4 raw files

m/z 400~525
54,624 MS/MS

m/z 525~650
53,705 MS/MS

m/z 650~775
52,974 MS/MS

m/z 775~900
52,562 MS/MS

36,734 HCD MS/MS spectra
**Running Time (Single-Threaded)**

_hours_ Hours

<table>
<thead>
<tr>
<th>400-525</th>
<th>525-650</th>
<th>650-775</th>
<th>775-900</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="green" alt="Decoy" /></td>
<td><img src="red" alt="Target" /></td>
<td><img src="blue" alt="LcMsMap" /></td>
<td><img src="green" alt="Decoy" /></td>
</tr>
<tr>
<td><img src="red" alt="Target" /></td>
<td><img src="green" alt="Decoy" /></td>
<td><img src="red" alt="Target" /></td>
<td><img src="green" alt="Decoy" /></td>
</tr>
<tr>
<td><img src="blue" alt="LcMsMap" /></td>
<td><img src="green" alt="Decoy" /></td>
<td><img src="red" alt="Target" /></td>
<td><img src="green" alt="Decoy" /></td>
</tr>
<tr>
<td><img src="green" alt="Decoy" /></td>
<td><img src="red" alt="Target" /></td>
<td><img src="blue" alt="LcMsMap" /></td>
<td><img src="green" alt="Decoy" /></td>
</tr>
</tbody>
</table>

**#Peptides: 4.2M** (target+decoy, fully tryptic, no modification)

Database: Target/Reversed Uniprot Mouse + Contaminants (16,835 proteins)

Core i7 3770K / 16GB / Single-Threaded
Running Time (PTM search)

Hours

<table>
<thead>
<tr>
<th>Hours</th>
<th>NoMod</th>
<th>Mod</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-525</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>525-650</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>650-775</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>775-900</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Power Outage!

Acetyl Protein N-Term, Oxidation M, Pyro-Glu N-term Q, Deamidation NQ
Up to 3 PTMs per peptide
Peptide Identifications (1% PepFDR)

IPA: A prototype implementation using MS-GF+; Very very very slow
QC-Shew from Q-Exactive

MS-GF+  MSPathFinder

1552  6944  1098

#PSMs

#Peptides

Bar chart showing the number of PSMs and peptides for MS-GF+ and MSPathFinder.
MSPathFinder is better than MS-GF+

1. It can find ultramodified peptides
   - To do: penalty for modifications

2. Does not require preprocessing
   - Input: raw file (currently only Thermo)
   - No need to run msconvert, precursor refinement tool
   - Works for both DDA and DIA
Also tested for data with 10 and 20 m/z isolation windows and it works!
Acknowledgements

PNNL
Christopher Wilkins
Grant Fujimoto
Samuel Payne
Matt Monroe
Tu Jin Shi
Weijun Qian
Tao Liu
Josh Adkins
Dick Smith

University of Washington
Michael MacCoss
Ying Sonia Ting
Jarrett Egertson
William S. Noble
KTH, Sweden
Lukas Käll