Comprehensive Peptide Searching Workflow to Maximize Protein Identifications

*Amol Prakash,[1] Shadab Ahmad,[1] David Sarracino,[1] Bryan Krastins,[1] MingMing Ning,[2]
Barbara Frewen,[1] Scott Peterman,[1] Gregory Byram,[1] Maryann S. Vogelsang,[1] Gouri Vadali,[1]
Jennifer Sutton,[1] Mary F. Lopez[1]
[1]Thermo Fisher Scientific, BRIMS (Biomarker Research in Mass Spectrometry), Cambridge, MA;
[2]Massachusetts General Hospital, Boston, MA*

Overview

Purpose: Development of a comprehensive protein identification workflow to maximize high-confidence peptide/protein identifications including post-translational modifications (PTM) compared to a traditional database search strategy.

Methods: Use of a combination of multiple search engines (e.g., SEQUEST®, Sequest HT, Mascot and MS Amanda) where combinations of PTMs were judiciously chosen for each node based on uniprotKB relative PTM abundances from high quality, manually curated, proteome-wide data[1].

Results: Tremendous enhancement in the high-confidence, Percolator-validated peptide and protein identifications compared to a standard protein identification workflow.

Introduction

Protein identification and characterization by mass spectrometry has become an established method in biological research in recent years. The number of protein identifications from complex biological samples depends on many factors, ranging from data acquisition strategy to MS/MS data searching methods. Unfortunately, only a fraction of spectra generated by the acquisition have confident peptide matches for any complex biological sample. There are several factors that are being overlooked by many users in the conventional data searching strategy, including the appropriate combination of PTMs, coding SNPs[2], isoforms of proteins, and iterative searching strategies that can potentially help to identify unmatched spectra. We developed a comprehensive MS/MS searching workflow in Thermo Scientific™ Proteome Discoverer™ software to maximize high-confidence peptide/protein identifications. The effect of various search strategy factors on peptide identifications were explored. We implemented a process that includes analysis of protein isoforms, missed cleavage sites, semi-tryptic digestion and most importantly, appropriate combination of PTMs in each search node. The workflows were tested on plasma and urine samples analyzed on a Thermo Scientific™ Orbitrap™ hybrid mass spectrometer. The comprehensive workflow was found to make more high-confidence peptide/protein IDs and identify multiple PTMs and partially cleaved peptides in a single run.

Methods

Comprehensive Workflow Development

We developed a comprehensive MS/MS searching workflow in Proteome Discoverer software using a combination of multiple search engines (Figure 1) in an iterative fashion to maximize protein/peptide identifications by considering the most frequently found PTMs1, artefacts (Table 1) and partially cleaved peptides. The combination of PTMs were judiciously chosen based on relative abundances (UniProtKB) of each PTM found experimentally and putatively as described in, from high-quality, manually curated, proteome-wide data1. The workflows were tested on plasma and urine samples analyzed on a hybrid Orbitrap mass spectrometer.

Sample Preparation

In order to evaluate the performance of the comprehensive workflow we took four human samples from two different sources (a) urine and (b) plasma (three samples). Human urine and plasma samples were collected with full consent and approval. The samples were subjected to reduction and alkylation followed by digestion with trypsin.

Liquid Chromatography and Mass Spectrometry

The digested samples were separated with a 5-45% acetonitrile gradient in 0.1% formic acid using a C18 nano-LC column. The urine sample (sample no. 1) and a plasma sample (sample no. 2) were run for 140 minutes and 90 minutes, respectively and the data were acquired with a Thermo Scientific™ LTQ Orbitrap Velos™ MS with Top 11 and Top 10 data-dependent MS/MS respectively, using CID fragmentation. Another two plasma samples (sample nos. 3 and 4) were run for 240 minutes and the data were acquired with the Thermo Scientific™ Q Exactive™ benchtop mass spectrometer, with Top 15 data-dependent MS/MS using HCD fragmentation.

Data Analysis

The acquired data was searched with Proteome Discoverer 1.4 against Uniprot human complete proteome database using the comprehensive workflow (Figure 1, Table1) and compared with the SEQUEST workflow with standard modifications (oxidation at methionine as dynamic modification and alkylation as static modification) coupled with percolator validation (Standard Search).

FIGURE 1. Structure of the comprehensive workflow

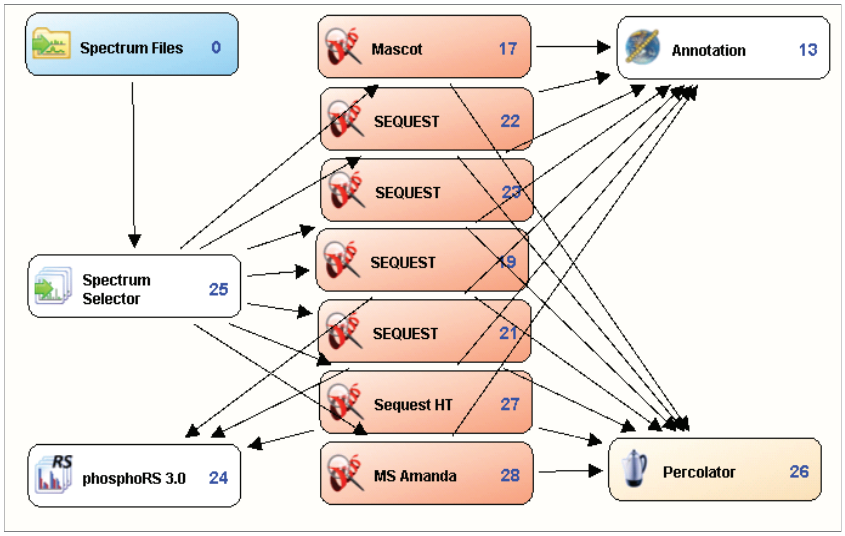

TABLE 1. Parameters and modifications used in comprehensive search workflow

Search Engine	Precursor Mass Tolerance	Fragment Mass Tolerance (Q Exactive MS/LTQ Orbitrap Velos MS)	Missed Cleavage	Enzyme	Static Modification	Dynamic Modification
Mascot	5 ppm	0.02 Da / 0.4 Da	2	Semi Trypsin	Carboxymethyl (C)	Oxidation (M); Acetyl (K); Methyl (K)
SEQUEST	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); ADP-Ribosyl (N,R); Myristoyl (K); Deamidation (N,Q); Phospho (S)
SEQUEST	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); Dioxidation (M); Trimethyl (K,R); Phospho (S,T)
SEQUEST	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); Carbamyl (K,R); Deamidated (N,Q); Amidation (Any C-Terminus)
SEQUEST	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); Methyl (K,R); Dimethyl (K,R); Trimethyl (K,R); Acetyl (K)
Sequest HT	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); Phospho (S,T,Y); Deamidated (N,Q);
MS Amenda	5 ppm	0.02 Da / 0.4 Da	3	Trypsin (Full)	Carboxymethyl (C)	Oxidation (M); Acetyl (K)

Results

We compared the results from our comprehensive searching strategy with a standard search strategy. We found that on average, the number of high-confidence peptide identifications (FDR≤0.01) increased approximately 2-fold with our comprehensive workflow compared to standard searches, whereas the increment in the number of medium confidence peptide identifications (FDR≤0.05) was more than two times compared to standard search (Figure 2).

The comprehensive workflow was found to increase the number of high-confidence protein identifications (FDR≤0.01) by 90% and the high-confidence protein groups by 75% with respect to the standard search condition. Moreover, the comprehensive workflow increases the high-confidence group proteins (with at least two high-confidence peptides for every protein in the group) by 23% (Figure 3).

The comprehensive workflow identified several high-confidence peptides with multiple PTMs which reveals the importance of particular combinations of PTMs in a search node (Table 2).

FIGURE 2. Comprehensive workflow increases number of peptide identifications (sample 1 = urine, sample 2-4 = plasma)

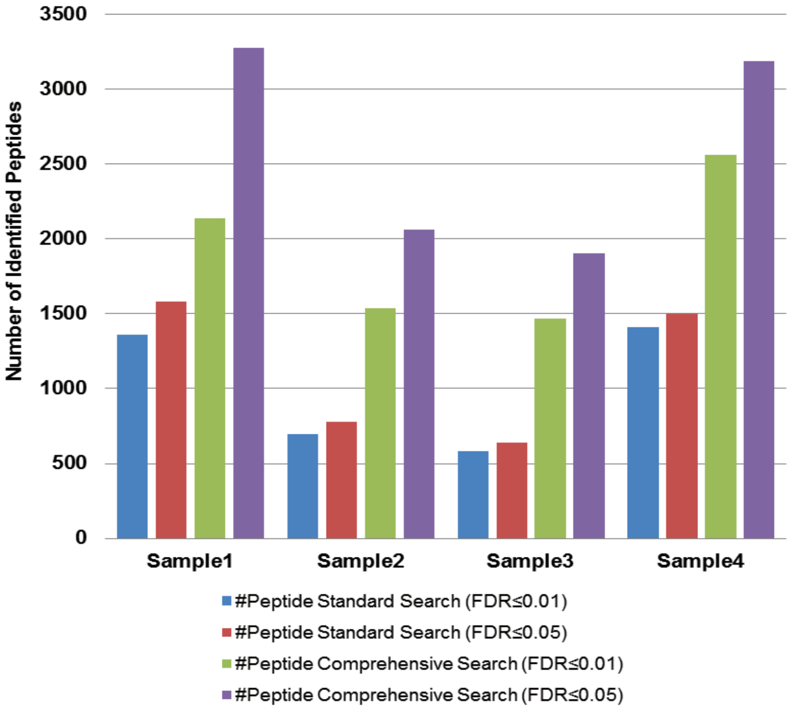

FIGURE 3. The comprehensive workflow increases the number of identified protein groups with at least two peptide hits per protein.

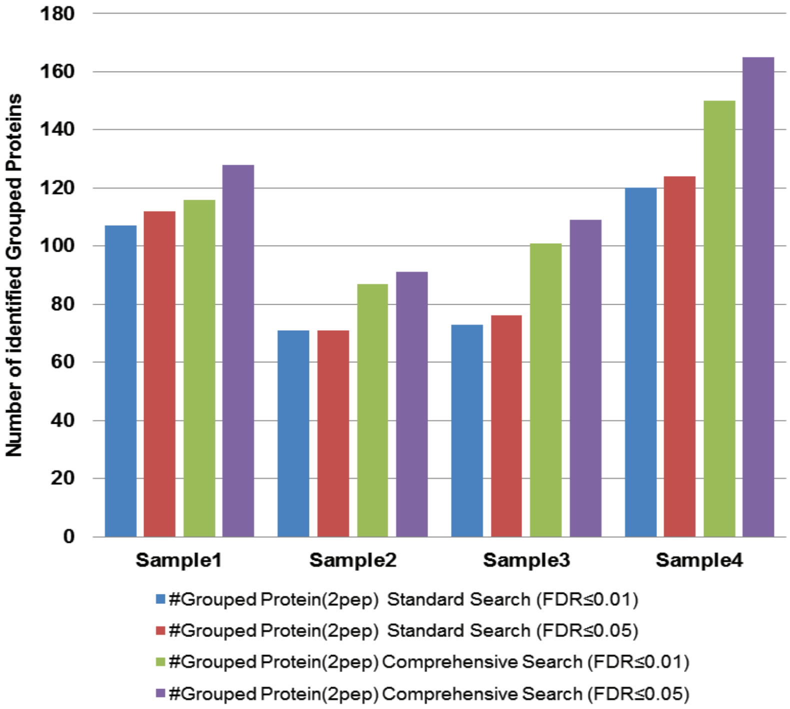

TABLE 2. Examples of peptides containing multiple PTMs from the comprehensive search strategy

Sequence	Modification	q-Value
CCKHPEAKRMPCAEDYLSVVLNQLCVLHEK	C1(Carboxymethyl); C2(Carboxymethyl); K3(Myristoyl); M10(Oxidation); C12(Carboxymethyl); C25(Carboxymethyl)	≤0.001
CYAKVFDEFKPLVEEPQNLIK	C1(Carboxymethyl); K4(Methyl); K10(Acetyl); K21(Methyl)	≤0.001
DKDEAEQAVSR	K2(Acetyl); R11(Trimethyl)	0.008
LVRPEVDVMCTAFHDNEETFLKK	R3(Dimethyl); M9(Oxidation); C10(Carboxymethyl); K22(Acetyl); K23(Acetyl)	0.004
INNEDNSQFK	N3(ADP-Ribosyl); K10(Myristoyl)	0.01
RMPCAEDYLSVVLNQLCVLHEK	R1(Trimethyl); M2(Dioxidation); C4(Carboxymethyl); C17(Carboxymethyl)	≤0.001
SEPKWEVVEPLK	K4(Trimethyl); K12(Dimethyl)	0.004
TCVADESAENCDK	C2(Carboxymethyl); C11(Carboxymethyl); K13(Dimethyl)	≤0.001
YYFNCNNWLSKVEGDRQWCR	C5(Carboxymethyl); K11(Methyl); R16(Trimethyl); C19(Carboxymethyl); R20(Methyl)	0.006

We further investigate the number of matched and unmatched spectra in the data sets comparing the standard search and our comprehensive search strategy. We found that the percentage of matched spectra improves significantly when using the comprehensive search workflow (Table 3).

Table 3. Comparative table for matched spectra

File	Total Spectra	Matched Spectra Standard Search (FDR≤0.05)	Matched Spectra Comprehensive Search (FDR≤0.05)	Matched Spectra Standard Search (FDR≤0.01)	Matched Spectra Comprehensive Search (FDR≤0.01)
Sample1	27215	28.0%	46.7%	26.2%	41.1%
Sample2	14005	15.4%	44.2%	14.4%	39.6%
Sample3	43036	5.1%	13.6%	4.9%	12.1%
Sample4	44450	9.5%	22.3%	9.0%	20.3%

Moreover, the comprehensive search workflow increased sequence coverage of proteins significantly, giving rich information about proteins including PTMs (Table 4).

Table 4. Comprehensive search increases protein coverage

Example	Protein	Sequence Coverage Standard Search (FDR≤0.01)	Sequence Coverage Comprehensive Search (FDR≤0.01)
1	A1AT	28.47%	57.42%
2	ALBU	70.94%	78.00%
3	A2MG	35.35%	53.12%
4	AACT	35.7%	42.55%
5	APOB	14.66%	23.12%
6	CERU	22.44%	37.28%
7	HEMO	38.96%	49.13%
8	TRFE	40.11%	61.17%
9	TTHY	54.42%	62.59%
10	VTDB	31.65%	50.21%

Conclusion

- A comprehensive workflow strategy identified almost twice as many high-confidence peptides compared to the standard search strategy.
- The comprehensive workflow helped increase the number of high-confidence protein identifications and high-confidence protein group identifications by approximately 90% and 75%, respectively, compared to the standard search approach.
- The comprehensive workflow identifies more high-confidence peptides with multiple PTMs.
- The percentage of matched spectra improves significantly when using the comprehensive search workflow in Proteome Discoverer software.

References

1. Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep. 2011 Sep 13;1.

2. Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, Andersen JS, Mann M. A mass spectrometry-friendly database for cSNP identification. Nat Methods. 2007 Jun; 4(6):465-6.

www.thermoscientific.com

Certified System™
ISO 9001
QMI-SAI Global

Thermo Fisher Scientific,
San Jose, CA USA
is ISO 9001:2008 Certified.

Africa-Other	+27 11 570 1840
Australia	+61 3 9757 4300
Austria	+43 810 282 206
Belgium	+32 53 73 42 41
Canada	+1 800 530 8447
China	+86 10 8419 3588
Denmark	+45 70 23 62 60
Europe-Other	+43 1 333 50 34 0
Finland	+358 9 3291 0200
France	+33 1 60 92 48 00
Germany	+49 6103 408 1014
India	+91 22 6742 9494
Italy	+39 02 950 591
Japan	+81 45 453 9100
Latin America	+1 561 688 8700
Middle East	+43 1 333 50 34 0
Netherlands	+31 76 579 55 55
New Zealand	+64 9 980 6700
Norway	+46 8 556 468 00
Russia/CIS	+43 1 333 50 34 0
South Africa	+27 11 570 1840
Spain	+34 914 845 965
Sweden	+46 8 556 468 00
Switzerland	+41 61 716 77 00
UK	+44 1442 233555
USA	+1 800 532 4752

Thermo
SCIENTIFIC
Part of Thermo Fisher Scientific