Electronic Supplementary Material (ESI) for Analyst.

Supporting Information

An Improved Scoring Method for the Identification of Endogenous Peptides based on Mascot MS/MS Ion Search

Ying-Lan Chen, Wei-Hung Chang, Chi-Ying Lee and Yet-Ran Chen*

Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan 11529

***To whom all correspondence should be addressed:** Telephone: +886-2-27872050, E-Mail: yetran@gate.sinica.edu.tw

Keywords: *Endogenous Peptide, Mass Spectrometry, Mascot Database Search and Plant Peptidomics*

Experimental Sections for Supporting Information

Liquid Chromatography-Tandem Mass Spectrometry Analysis of Tryptic Peptides from MassPREP Digestion Standard Mix 1

The tryptic peptides of MPDS Mix 1 were analyzed by a nanoUHPLC system (nanoACQUITY UPLC, Waters, Millford, MA) coupled with the nanoelectrospray ionization (nESI) source of a hybrid quadrupole time-of-flight mass spectrometer (Q-TOF-MS) (SYNAPT HDMS G2, Waters, Manchester, UK). In the LC-MS/MS analysis, water with 0.1% FA and ACN with 0.1% FA were used as the mobile phase. The sample was injected into a 180 μm × 50 mm tunnel frit trap column packed with 20 mm × 5 μm particles (Symmetry C18, Waters, Milford, MA) and separated online with a reverse phase analytical column (BEH C18, 1.7 μm, 75 μm × 250 mm, Waters, Milford, MA) at the flow rate of 300 nl/min using a 95 min gradient with 5-90% ACN ratio.[1] The mass spectrometry was operated in the positive ion mode and data dependent acquisition (DDA) methods were applied. The DDA settings were set to one full MS scan (400-1600 m/z) with a scan time of 0.6 second and switched to 1 product ion scans (50-1900 m/z) with a 0.6 second scan time when a precursor ion charge was 2+, 3+ or 4+ and the intensity was higher than 1500 counts.

Spectra Processing

The SYNAPT HDMS G2 data generated for the analysis of tryptic peptides in MPDS Mix 1 were converted into mzXML format using massWolf (version 4.3.1) and processed by UniQua (version 1.0). The UniQua parameters for the processing the MS/MS spectra from SYNAPT HDMS G2 were smoothing = 7, centroiding high = 80%, maximum resolution = 25000, baseline cutoff = 30 counts. The SYNAPT HDMS G1 data generated for the analysis of the endogenous

peptides from wounded tomato in our previous study were converted into mzXML format[2] using massWolf (version 4.3.1) and then processed by UniQua[3] (version 1.0).[2-4] The UniQua parameters for processing the MS/MS spectra from SYNAPT HDMS G1 were smoothing = 9, centroiding high = 80%, maximum resolution = 16000, baseline cutoff = 1.5 counts. All processed MS/MS spectra were extracted and converted into Mascot generic format (.mgf) using mzXML2Search in Trans Proteomics Pipeline (TPP) version 4.4 rev. 1. The downloaded LTQ-velos Orbitrap datasets were converted into mgf files using MSconvert of ProteoWizard.[5]

Peptide Identification Database

For identification of the peptides in MPDS Mix 1, a concatenated protein database containing forward and randomized sequences of NCBI *Escherichia coli* (*E. Coli*) O127:H6 str. E2348/69 RefSeq Genome database with the protein sequence in MPDS Mix 1 (total protein entries, 9,314; *E. coli* protein sequences with 4 protein sequences of yeast enolase 1, bovine serum albumin, rabbit glycogen phosphorylase and yeast alcohol dehydrogenase 1) was used. For identification of the peptides from yeast cell lysate spiked with different concentrations of UPS1 standard mixture, a concatenated protein database containing forward and randomized sequences of Ensembl *Saccharomyces cerevisiae* (*S. cerevisiae*) R64-1-1 Genome database with UPS1 standard protein mixture (total protein entries, 13,296; *S. cerevisiae* protein sequences with 48 sequences of UPS1 human recombinant proteins) was used. For identification of tomato endogenous peptides, a concatenated protein database containing forward and randomized sequences of tomato protein sequences from the International Tomato Annotation Group (ITAG) protein database version 2.3 (total protein entries, 69,456; with the addition of bovine β-casein sequence) was used. The protein sequence randomization was performed using the Perl script (decoy.pl) tool provided by Matrix Science (London, UK).

Reference

1. Chen, C. J.; Chen, W. Y.; Tseng, M. C.; Chen, Y. R., Tunnel frit: a nonmetallic in-capillary frit for nanoflow ultra high-performance liquid chromatography-mass spectrometryapplications. *Anal Chem* **2012,** *84* (1), 297-303.
2. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology* **2004,** *22* (11), 1459-66.
3. Chang, W. H.; Lee, C. Y.; Lin, C. Y.; Chen, W. Y.; Chen, M. C.; Tzou, W. S.; Chen, Y. R., UniQua: a universal signal processor for MS-based qualitative and quantitative proteomics applications. *Anal Chem* **2013,** *85* (2), 890-7.
4. Chen, Y. L.; Lee, C. Y.; Cheng, K. T.; Chang, W. H.; Huang, R. N.; Nam, H. G.; Chen, Y. R., Quantitative Peptidomics Study Reveals That a Wound-Induced Peptide from PR-1 Regulates Immune Signaling in Tomato. *Plant Cell* **2014,** *26* (10), 4135-4148.
5. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **2012,** *30* (10), 918-20.

Figure Legend

Table S-1. List of the 8 peptides identified by Mascot score but excluded by DS

Table S-2. List of the MPDS Mix1 peptide hits confidently identified by Mascot score, DS and CS using NES and sTPS searches based on FDR < 0.01

Table S-3. List of the 7 and 5 peptides identified by CS but excluded by DS in NES search and in sTPS search

Table S-4. List of the tomato endogenous peptide hits confidently identified by Mascot score, DS and CS based on FDR < 0.01

Table S-5. List of all target and decoy hits with q-value for tomato endogenous peptide analysis using Mascot Percolator

Figure S-1. Score distributions of NES and sTPS searched (A) forward and (B) random matched hits using Mascot NES and sTPS search against a concatenated database containing forward and randomized sequences of *E.coli* and MPDS Mix 1 to identify the tryptic peptides of MPDS Mix 1. The black dash line indicated the score threshold for 1% random match probability evaluated by Mascot.

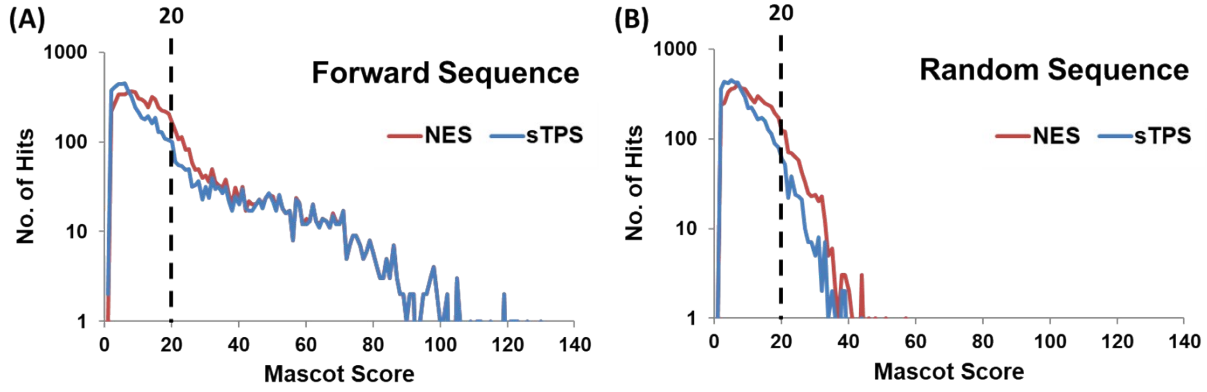
Figure S-2. Score distributions of forward and random matched hits using Mascot NES search against a concatenated database containing forward and randomized sequences of *E.coli* and MPDS Mix 1 to identify the tryptic peptides of MPDS Mix 1 for each Mascot sub-rank (the 1^{st} to 10^{th} rank).

Figure S-3. Precision versus sensitivity curves for the identification of the tryptic peptides of MPDS Mix 1 using sTPS search with different scoring methods. Sensitivity is the coverage of total number of the spectra matched to the correct peptide sequences. Precision (Correct Rate) is the ratio for the number of correct and total peptide hits. The line indicates the 99% correct rate.

Figure S-4. (A) Mass deviations of the observed precursor mass and (B) score distributions of correct (1263 hits for forward sequences of 4 standard proteins, denoted as red open circle) and incorrect (11849 hits for random sequences of 4 standard proteins + forward and random sequences of *E. coli* proteins, denoted as blue diamond) hits using NES search and Mascot score, delta score (DS) and contribution score (CS)

for the identification of the tryptic peptides of MPDS Mix 1. The dashed line shows the cut-off score threshold for FDR < 0.01.

Figure S-5. Numbers of the identified nonredundant UPS1 peptides in the analysis of the tryptic peptides from yeast proteins spiked with different concentrations (50, 25 and 0.5 fmol) of UPS1 using NES search with Mascot score, DS and DS + CS.



	Area of NES	Area of sTPS	NES/ sTPS Ratio
Hits with Mascot score >20	69875.4	59480.4	1.17

	Area of NES	Area of sTPS	NES/ sTPS Ratio
Hits with Mascot score >20	15957.9	6524.2	2.45

Figure S-1. Score distributions of NES and sTPS searched (A) forward and (B) random matched hits using Mascot NES and sTPS search against a concatenated database containing forward and randomized sequences of *E.coli* and MPDS Mix 1 to identify the tryptic peptides of MPDS Mix 1. The black dash line indicated the score threshold for 1% random match probability evaluated by Mascot.

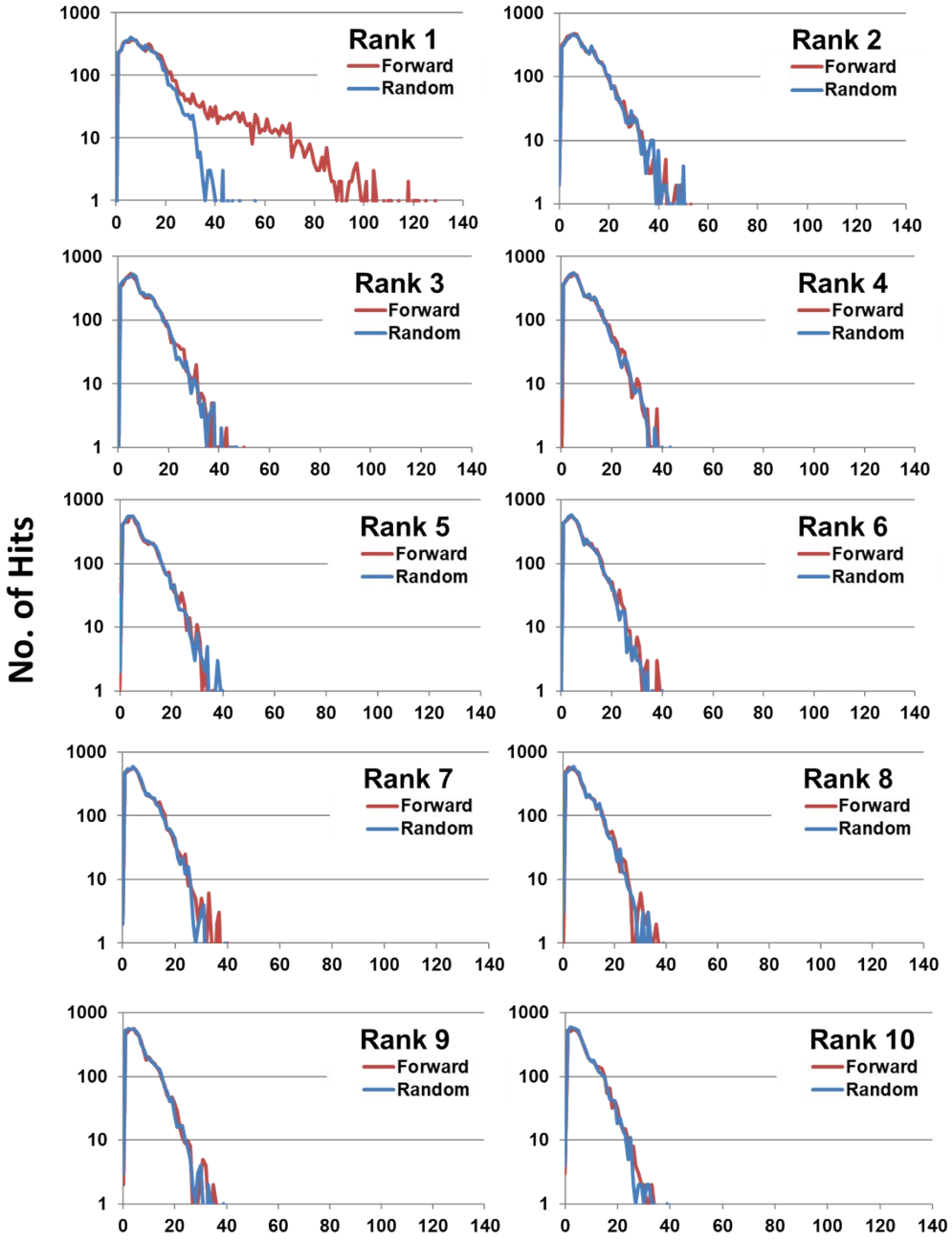


Figure S-2. Score distributions of forward and random matched hits using Mascot NES search against a concatenated database containing forward and randomized sequences of *E.coli* and MPDS Mix 1 to identify the tryptic peptides of MPDS Mix 1 for each Mascot sub-rank (the 1st to 10th rank).

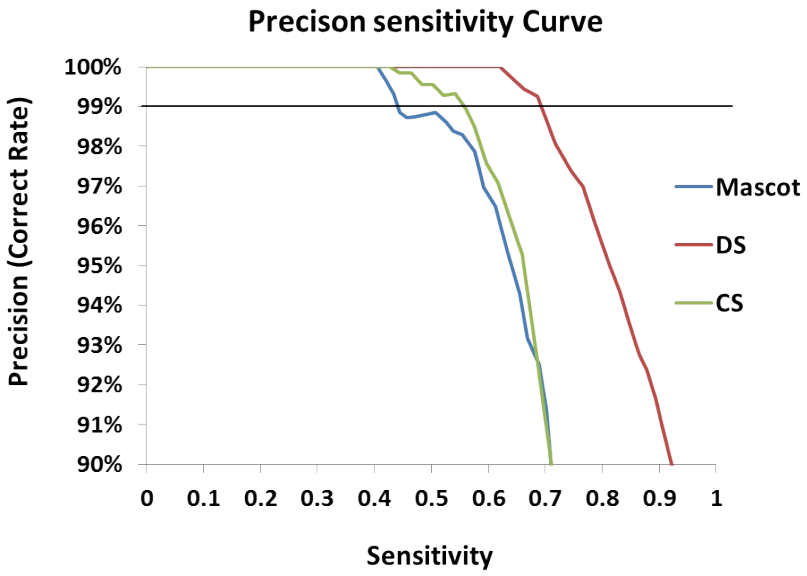


Figure S-3. Precision versus sensitivity curves for the identification of the tryptic peptides of MPDS Mix 1 using sTPS search with different scoring methods. Sensitivity is the coverage of total number of the spectra matched to the correct peptide sequences. Precision (Correct Rate) is the ratio for the number of correct and total peptide hits. The line indicates the 99% correct rate.

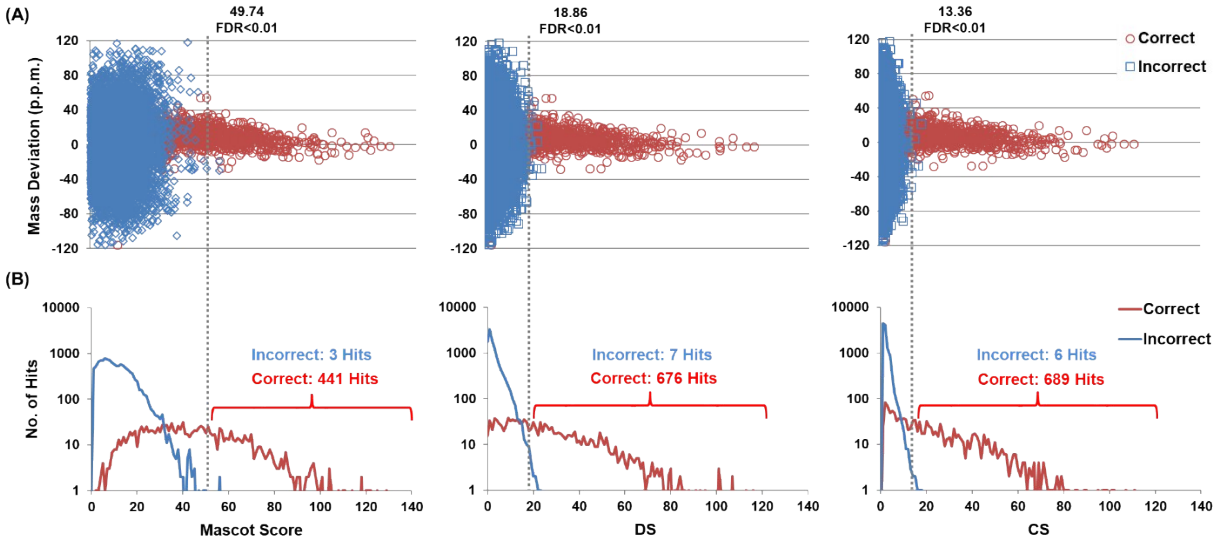


Figure S-4. (A) Mass deviations of the observed precursor mass and (B) score distributions of correct (1263 hits for forward sequences of 4 standard proteins, denoted as red open circle) and incorrect (11849 hits for random sequences of 4 standard proteins + forward and random sequences of *E. coli* proteins, denoted as blue diamond) hits using NES search and Mascot score, DS and CS for the identification of the tryptic peptides of MPDS Mix 1. The dashed line shows the cut-off score threshold for FDR < 0.01.

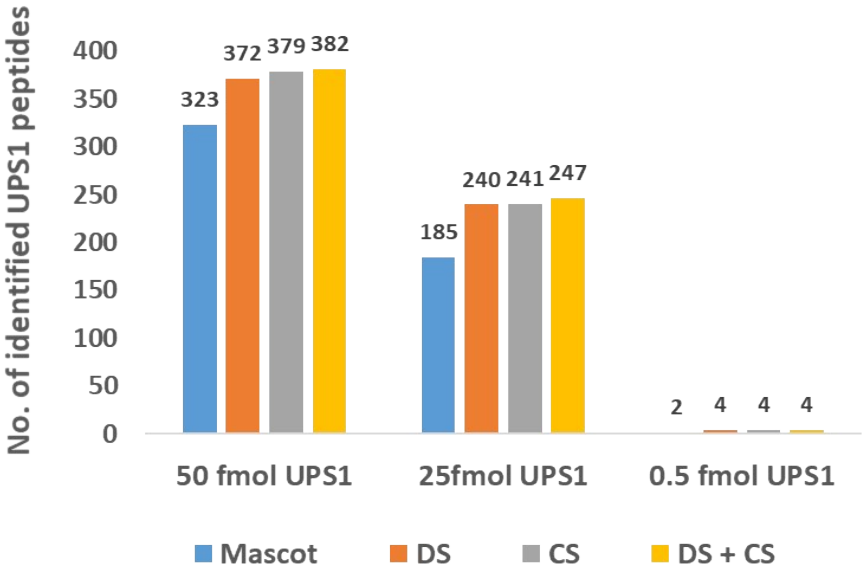


Figure S-5. Numbers of the identified nonredundant UPS1 peptides in the analysis of the tryptic peptides from yeast proteins spiked with different concentrations (50, 25 and 0.5 fmol) of UPS1 using NES search with Mascot score, DS, CS and DS + CS.