A Novel "Ion Accounting" Algorithm for Protein Database Searches

Guo-Zhong Li, Dan Golick, Marc V. Gorenstein, Johannes P.C Vissers, Jeffrey C. Silva, Scott J. Geromanos
Waters Corporation, Milford, MA

OVERVIEW

- A novel database search algorithm is described that is ideally suited for identifying proteins over a wide dynamic range.
- Alternate scanning LC-MS (LC-MS^E)[1,2] data from a nanoACQUITY UPLC™ chromatograph coupled with a Q-Tof Premier™ mass spectrometer is utilized in this search algorithm (Fig.1).
- The database search strategy ("Ion Accounting") is a hierarchal, protein-centric search algorithm containing three incrementally stringent modules (Fig. 10).
  1. Raw Tryptic Peptide/Protein Identification: Time-resolved accurate mass, LC-MS^E data is matched directly to a protein database to produce multiple, tentative peptide identifications.
  2. Stringent Tryptic Peptide/Protein Identification: Stringent search criteria mass and intensity-based peptide/protein attributes are used to score the peptide/protein identifications, above a user-specified false identification rate.
  3. Expanded Tryptic Peptide/Protein Identification: A subset data base search of the identified proteins is performed to increase peptide sequence coverage of the identified proteins. The search criteria is also expanded to include corresponding in-source fragments, neutral losses (H_2O and NH_3), missed cleavages, and user-specified variable modifications (acetylation, phosphorylation, deamidation,...).

LC-MS^E ANALYSIS

Fig. 1

Fig. 1. LC-MS^E is an alternate-scanning technique, for data-independent acquisition of precursors and fragments.[1,2] During the MS acquisition, the Q- Tof collision cell is held at low potential for precursor ion detection, while in the MS^E cycle (multiplex fragmentation) the collision cell is at an elevated potential for fragment ion detections. The MS data provides, accurate peptide mass measurements that can be used for subsequent quantitation across multiple analyses. The time-resolved MS^E data represents fragmentation data for all detected precursors. The LC-MS^E methodology requires that precursors and associated fragments show identical chromatographic profiles (Fig. 2).

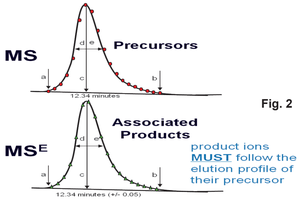


Fig. 2

DATA & DATABASE

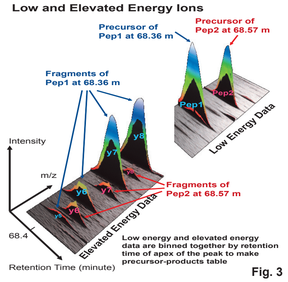


Fig. 3

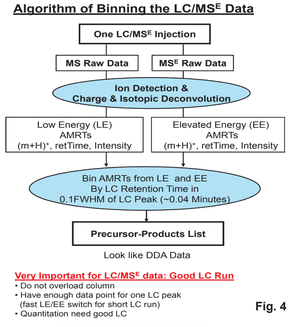


Fig. 4

Fig. 5

Fig. 3 and Fig. 4, Precursor-Product ion tables are produced for each detected precursor by time-aligning low (precursor) and elevated energy (product) ions within a retention time tolerance (~10% FWHM) of the chromatographic peak.
Fig. 5, Peptide-Fragment tables are produce *in silico* from protein FASTA database.

DATABASE SEARCH

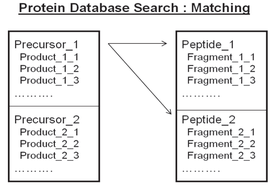


Fig. 6

Fig. 6. Each detected precursor can produce matches to multiple peptide sequences in the database at 10 ppm mass accuracy. The "Ion Accounting" search strategy dictates that *detected precursor and product mass measurements be used for the validation of a single peptide.* ***A single mass measurement can not be used to validate multiple peptides (Fig 10-11).***

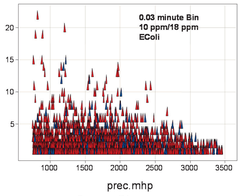


Fig. 7

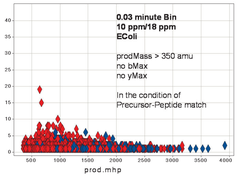


Fig. 8

Fig. 9 Rank Peptides and Proteins for Depletion

Protein ID	Peptide Sequence	Precursor	Peptide Rank	For Rank Protein
1	PepSeq_1_1	Precursor_1	1	yes
1	PepSeq_1_2	Precursor_2	1	yes
1	PepSeq_1_3	Precursor_3	1	yes
1	**PepSeq_1_4**	**Precursor_A**	**2**	**No**
1	PepSeq_1_5	Precursor_4	1	yes
2	PepSeq_2_1	Precursor_4	1	yes
2	PepSeq_2_2	Precursor_5	1	yes
2	**PepSeq_2_3**	**Precursor_A**	**1**	**yes**
2	PepSeq_2_4	Precursor_6	1	Yes
2	PepSeq_2_5	Precursor_B	2	No

SEARCH ALGORITHM

Fig. 10

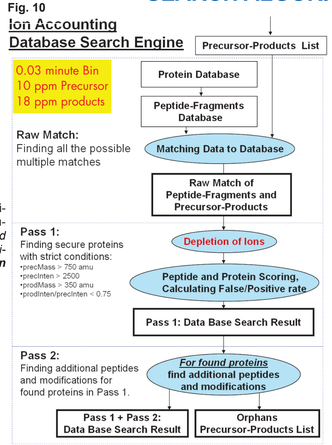
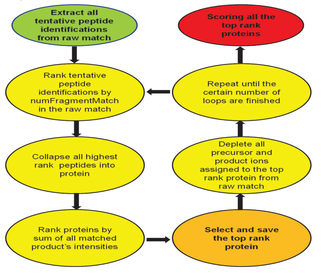


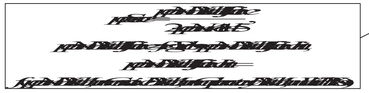
Fig. 11 Depletion of Ions in Pass 1



$$\text{proteinScore} = \text{protPhysicalProperty} \bullet \sum_i pepScore(i) \bullet \frac{2 \cdot pepNumAAcid - 5}{2 \cdot protNumAAcid - 5 \cdot protNumPep}$$

$$\text{protPhysicalProperty} = 0.67 \bullet \left(\frac{protNumBYMatch}{0.6 \cdot protTheorNumBYMatch} \right)^{1.5}$$

$$protTheorNumBYMatch = 3.0 + 0.003 \cdot \exp(1.12 \cdot \log 10(protNumAAcid * ProtSumPrecIntenMatch))$$



What is in the output:

Protein (score, false/positive rate, sumPrecInten, seqCover, ...)
Peptide (score, sumProdInten, ...),
precursor (mass, retT, inten),
byFragmentInfo, **product** (mass, retT, inten),
precMassError, prodMassError, precProdRetTBinError

How do we report result

After Pass 1, all the proteins are sorted by protein score. False/positive rates are calculated for each protein based on protein type (regular or Reverse/Random). Final protein cutoff is by false/positive rate from user or default (4.0%).

prot.dataBaseType	prot.id	prot.score	prot.falsePositiveRate
Regular	1199	770.38	0
Regular	2083	446.74	0
Regular	781	176.95	0
Regular	3271	0.21	0
.....	.....	.....	.....
Reverse	7225	0.2	0.52
Regular	858	0.19	0.52
Regular	115	0.18	0.51
Regular	1673	0.18	0.51
Regular	1615	0.17	0.51
Regular	717	0.16	0.5
Reverse	4811	0.16	1
Regular	660	0.16	1
Regular	1222	0.16	0.99
Regular	2338	0.16	0.99
Regular	4071	0.15	0.98
Regular	3711	0.15	0.97
Reverse	5115	0.15	1.44
.....	.....	.....	.....

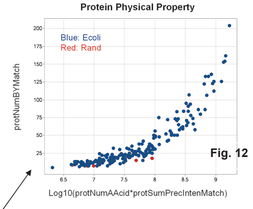


Fig. 12

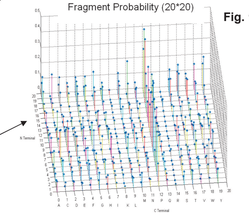


Fig. 13

RESULT

299 found proteins in EColi, 4% false positive rate with 1x additional random database

	Number identified peptides
Secure Peptides (Pass 1)	2201
Additional Peptides (Pass 2)	1073
Miss Cleave Peptides (Pass 2)	676
SpecialVarMod (Pass 2)	63
phosVarMod (Pass 2)	5
inSourceFrag (Pass 2)	1480
precNLoss (Pass 2)	288

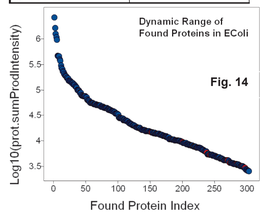


Fig. 14

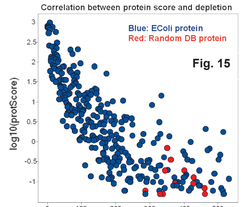


Fig. 15

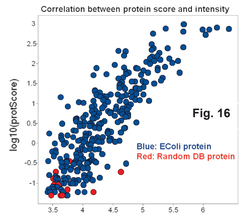


Fig. 16

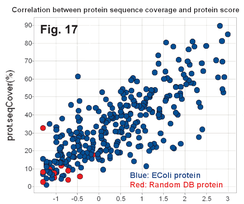


Fig. 17

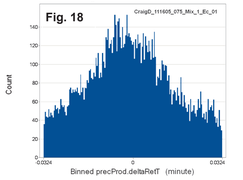


Fig. 18

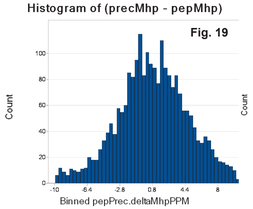


Fig. 19

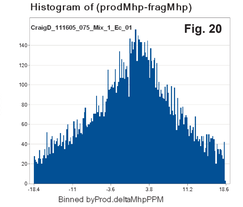


Fig. 20

CONCLUSIONS

- **A novel, hierarchal protein database search algorithm ("Ion Accounting") has been developed to accommodate multiplexed LC-MS analysis (LC-MS^E) of complex proteins.**
- **The algorithm relies on time-resolved (~0.03 min) mass measurements of precursors (<10ppm) and fragments (<20ppm) obtained from an LC-MS^E analysis. Precise time resolution enables efficient binning of fragments to each detected precursor.**
- **Protein and peptide scoring utilizes the information obtained from the quantitative acquisition of each detected precursor and fragment. These attributes include empirically derived peptide and protein physical properties. Additional rules such as consecutive fragmentation data and fragmentation probabilities are also integrated into the scoring of peptides and proteins (Fig. 12-16).**
- **The "Ion Accounting" algorithm dictates that each precursor and product ion can only be used once for any peptide identification.**

ACKNOWLEDGMENTS

Barry Dyson, Craig Dorschel , Beverly Kenney, Martha Stapels, Petra Olivova, Yingqing Yu, Asish Chakraborty, Hongji Liu, Weibein Chen, Martin Gilar, Scott Berger, James Langridge, Therese McKenna, Ian Campuzano, Phillip Young, Keith Richardson, Richard Denny

REFERENCES

1. Silva, et al. Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. ***Anal. Chem.*** **2005, 77, 2187 - 2200.**
2. Silva, et al. Absolute Quantification of Proteins by LC/MS^E. ***Mol. Cell. Proteomics*** **2005, 5, 145-156.**

TO DOWNLOAD A COPY OF THIS POSTER VISIT WWW.WATERS.COM/POSTERS

 720001921EN