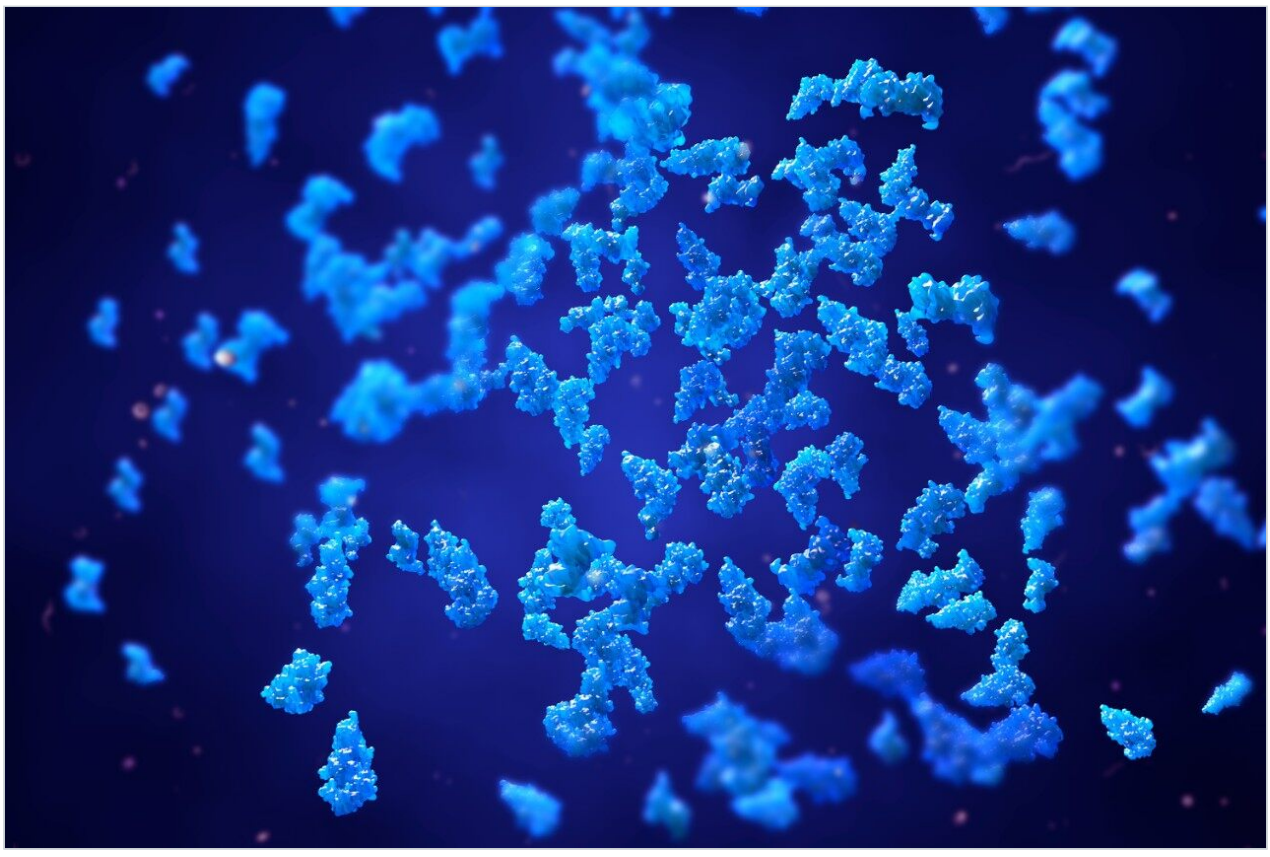
Waters™

Application Note

Identity^{E}: A Novel Database Search Strategy for Accurate Mass LC-MS^{E} Data

Guo-Zhong Li, Marc Gorenstein, Daniel Golick, Martha D. Stapels, Hans Vissers, Barry Dyson, James I. Langridge, Scott J. Geromanos

Waters Corporation



Abstract

- The false positive identification of proteins has been a major concern in many proteomic projects. As such, it is becoming unacceptable to report one peptide identification in peer reviewed publications.
- A new iterative databank searching approach is described that uses accurate mass data and other peptide properties, such as retention time, to remove the spurious identification of proteins.
- This strategy has been designed specifically to search data-independent, multiplexed LC-MS^E data, which provides increased peptide coverage over data dependant LC-MS/MS experiments.
- A reverse/random decoy database is automatically generated to determine the false positive rate.

Introduction

A concern of the proteomics community is the confidence of a peptide and protein assignment from tandem MS data. The number of single peptide protein identifications present in peer reviewed publications is unaccountably high, which raises doubts over the validity of the results.

LC-MS^E is a parallel unbiased approach to data acquisition that increases both the number of peptides and also the reproducibility of the peptides sampled during an LC-MS experiment. A novel database search algorithm is presented for the qualitative identification of data originating from LC-MS^E data, whereby multiple precursor ions are fragmented simultaneously. Properties that are used by the algorithm include retention-time, precursor and product ion intensities, charge states, and crucially the accurate masses of both the precursor and product ions from the LC-MS^E data. This strategy has been shown to be highly effective for the identification of proteins in both simple and complex samples over a wide dynamic range.

The database search algorithm is an iterative process whereby each iteration incrementally increases the selectivity, specificity, and sensitivity of the overall strategy.

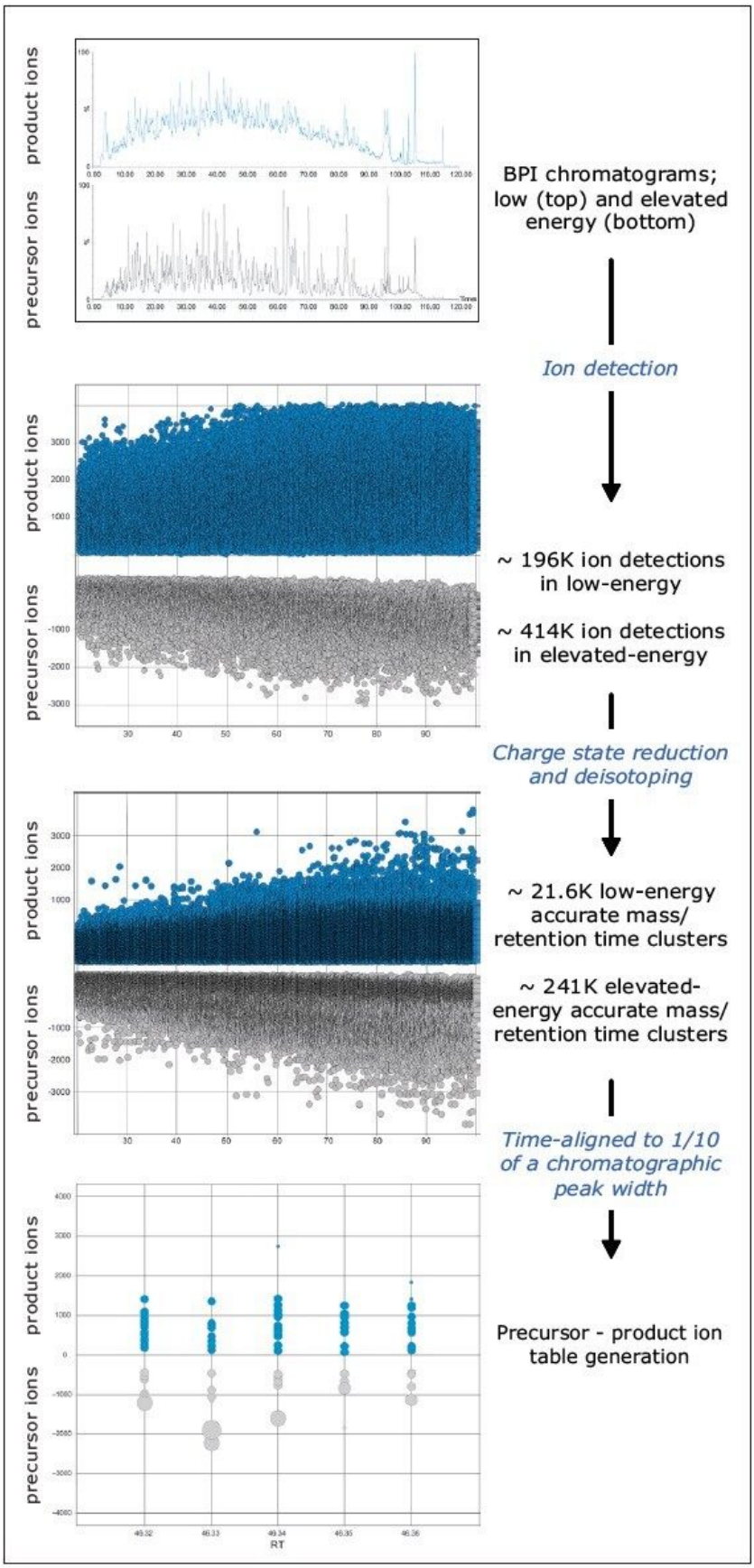


Figure 1. LC-MS^E

data processing, showing the ion detection, time-alignment, and deconvolution process.

Tentative peptide and protein identifications are ranked and scored by their relative correlation to a number of well-established models of known and empirically derived physicochemical attributes of proteins and peptides. The algorithm utilizes reverse or random decoy databases for automatically determining the false positive identification rate.

The data presented demonstrates the ability of the method to correctly identify peptides and proteins from data-independent acquisition strategies with high sensitivity and specificity.

Databank Searching

After data acquisition and processing, which generates a file containing precursor and product ion masses for each peptide, the user defines several parameters for databank searching e.g. database, database type, and whether to create a reverse or random decoy database (see Figures 2a to 2d).

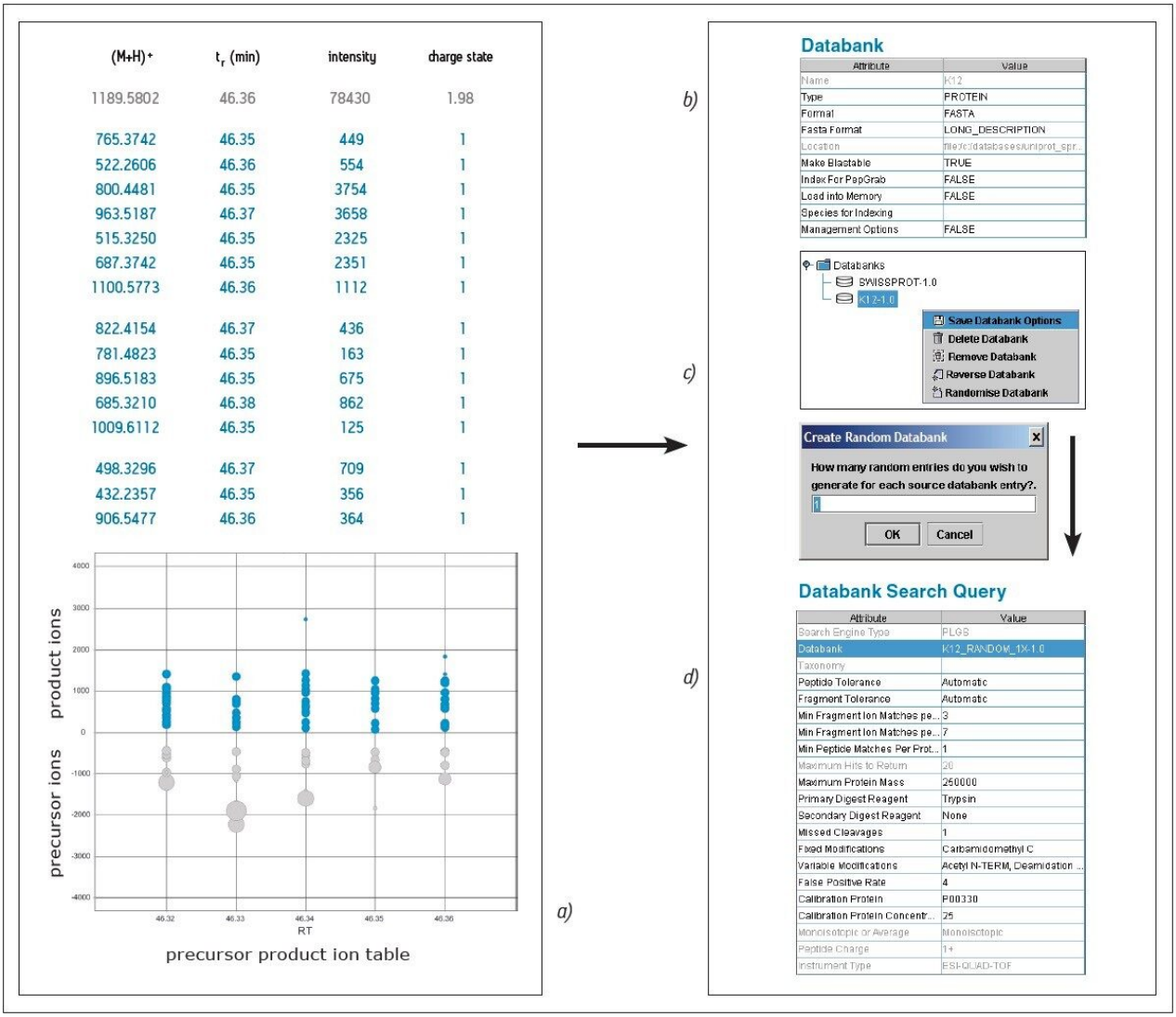


Figure 2. Defining the databank search parameters: a) Filtered precursor-product ion table, b) Database selection, c) Generating a decoy database, d) Setting up the database query parameters

Experimental

Materials and Methods

Sample

The sample used in this study was a mixture of four standard proteins, tryptically digested and spiked into a digested cytosolic lysate of *E. coli*. The total column load was 500 ng: 460 ng of the cytosolic *E. coli* lysate

and 40 ng of the four standard proteins.

LC System

A nanoACQUITY UPLC System was used with 75 µm x 10 cm bridged ethyl hybrid C_{18} (1.7 µm); Gradient: 3 - 40% B for 90 min @ 250 nL/min; eluent A and B: 0.1% formic acid in water and acetonitrile, respectively.

MS System

A SYNAPT MS System was operated in the LC-MS^E alternate scanning mode. Low and elevated energy spectra were acquired every 1.5 s; Collision energy ramp elevated energy: 15 - 40 volts over 1.5 s; Lock Mass: 100 fmol/µL [Glu1]-Fibrinopeptide B @ 500 nL/min sampled once every 30 s; TOF resolution: 10,000 (v mode of acquisition).

Database Search Workflow

A flow diagram showing the hierarchical principle of the database search algorithm is illustrated in Figure 3. A pre-assessment survey – to assess this particular experimental dataset – and a database search encompassing the physicochemical properties of peptides and proteins in the liquid and gas phase was conducted in a so-called “first pass” search. This process was followed by a peptide ranking process and collapsing the identified peptides into proteins.

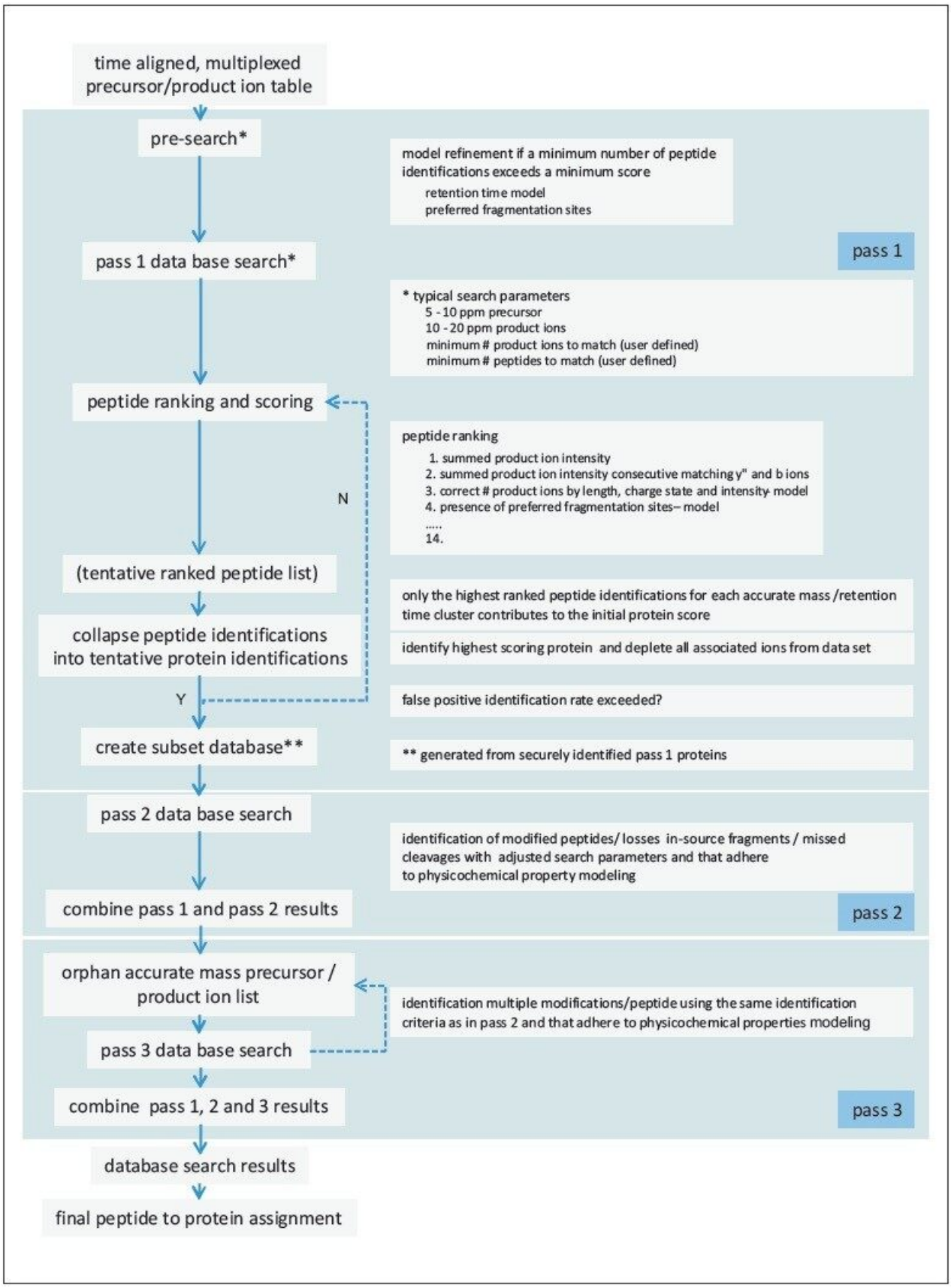

time aligned, multiplexed precursor/product ion table
pre-search*
model refinement if a minimum number of peptide identifications exceeds a minimum score
retention time model
preferred fragmentation sites
pass 1 data base search*
pass 1
* typical search parameters
5 - 10 ppm precursor
10 - 20 ppm product ions
minimum # product ions to match (user defined)
minimum # peptides to match (user defined)
peptide ranking and scoring
peptide ranking
1. summed product ion intensity
2. summed product ion intensity consecutive matching y" and b ions
3. correct # product ions by length, charge state and intensity- model
4. presence of preferred fragmentation sites– model
.....
14.
N
(tentative ranked peptide list)
only the highest ranked peptide identifications for each accurate mass /retention time cluster contributes to the initial protein score
collapse peptide identifications into tentative protein identifications
identify highest scoring protein and deplete all associated ions from data set
Y
false positive identification rate exceeded?
create subset database**
** generated from securely identified pass 1 proteins
pass 2 data base search
identification of modified peptides/ losses in-source fragments / missed cleavages with adjusted search parameters and that adhere to physicochemical property modeling
combine pass 1 and pass 2 results
pass 2
orphan accurate mass precursor / product ion list
identification multiple modifications/peptide using the same identification criteria as in pass 2 and that adhere to physicochemical properties modeling
pass 3 data base search
combine pass 1, 2 and 3 results
pass 3
database search results
final peptide to protein assignment

Figure 3. Overview of the ion accounting databank search process: a) Database search - encompassing modeling b) Peptide ranking c) Protein ranking

Following this, a subset database was generated and a second pass search conducted, which was subsequently used to identify user-defined variable peptide modifications and peptide fragments.

The results of the various search iterations were mapped together and a protein ranking process initiated.

The results from this iterative search program illustrate a high degree of replication at both the peptide and protein levels and this is presented in Figures 4 and 5, respectively. The latter addresses one of the major concerns for dealing with tandem MS/MS data in proteomics experiments.

Database Search Results- Replication Rate at the Peptide and Protein Level

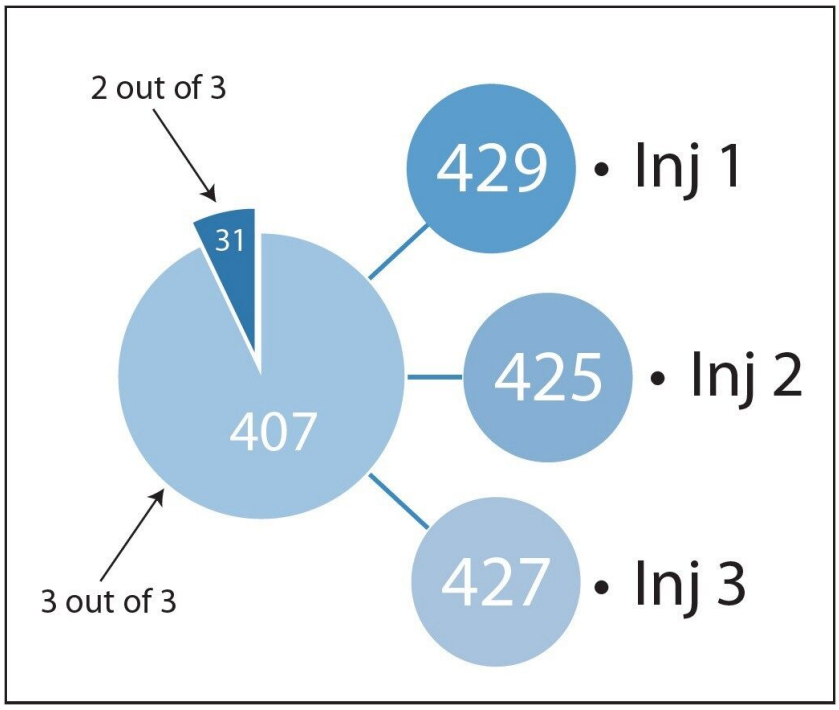


Figure 4. The number of proteins identified in each replicate injection of the E. coli sample (n=3) and the corresponding intersection. A total of 438 proteins replicate in 2 out of 3 injections. This number includes five standards and trypsin.

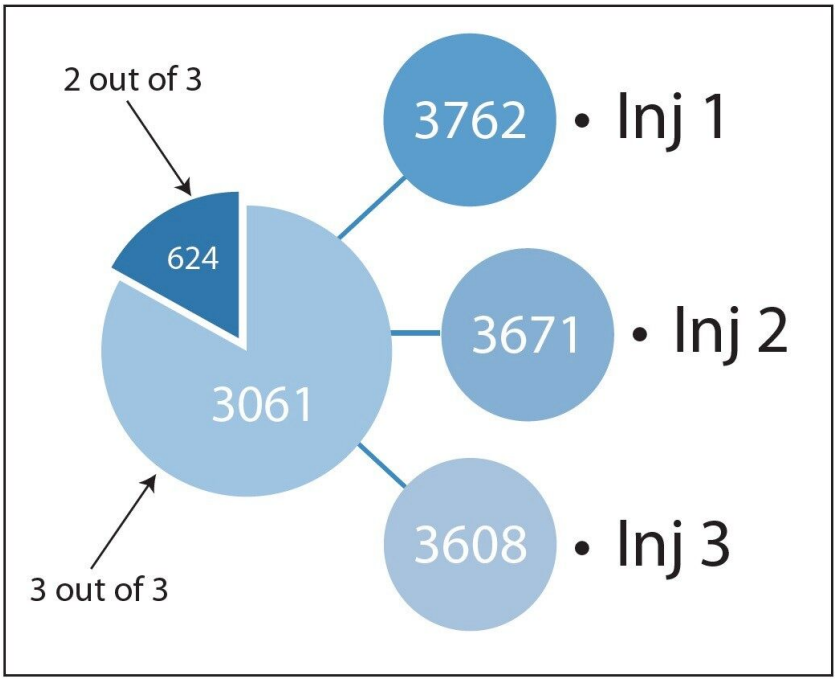


Figure 5. The number of peptides identified in each replicate injection of the E. coli sample (n=3) and the corresponding intersection.

Results and Discussion

Selectivity, Specificity, and Sensitivity

To illustrate the selectivity, specificity, and sensitivity of the scoring and validation process, an *E. coli* LC-MS^E dataset was queried against species-specific databases of six different bacterial proteomes, using a 4% acceptable false positive rate. Searching of the *E. coli* data against the different bacterial proteomes should only result in the identification of homologous proteins between the organisms, and not a large number of low scoring, spurious proteins. This is a common failing with traditional tandem MS data and existing search engines.

The results from the Identity^E search are displayed in Figure 6, with peptides displayed in magenta identified to a species-specific protein. It can clearly be seen that an amazing degree of specificity is afforded by the search strategy, providing confidence in the results obtained. The blue, red, and green ions represent matched ions corresponding to tryptic peptides, missed cleavage products, and variable modifications from YEAST_ADH, one of the exogenous spiked in proteins.

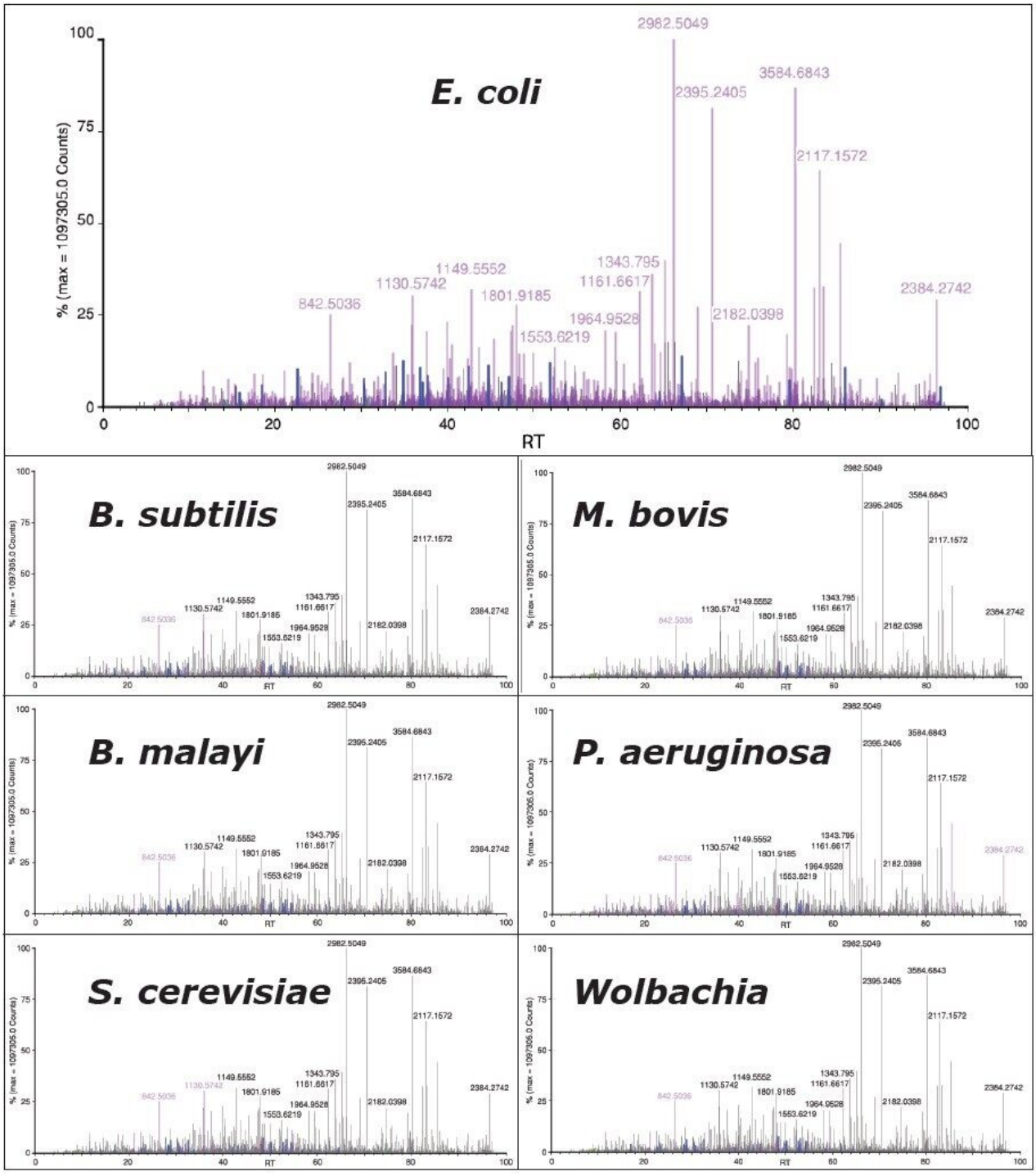


Figure 6. Results from an E. coli LC-MS^E dataset queried against six species specific databases from different bacterial proteomes. Grey = non-identified. The actual number of identified proteins is presented in Figure 7.

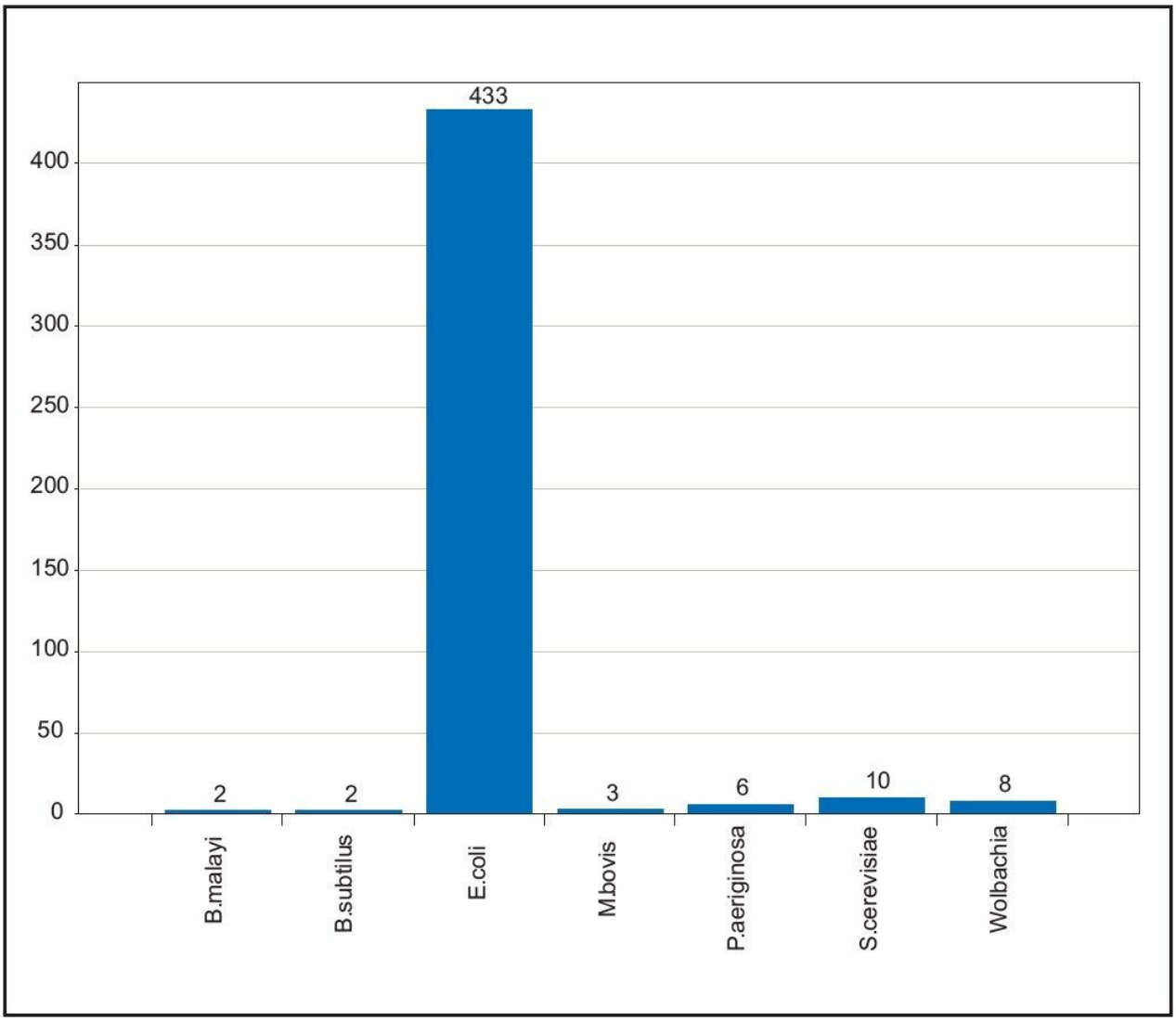


Figure 7. Search results illustrating the number of identified proteins from an E. coli LC-MS^E dataset searched against six different species specific databases from bacterial proteomes.

Absolute Quantification

An added unique benefit of the LC-MS^E data is the ability to generate absolute quantification values for each identified protein, that contains more than two peptides[1]. For the *E.coli* dataset this is illustrated in Figure 8, where absolute amounts for each one of the proteins is shown. It can be observed that close to three orders of magnitude of identification dynamic range can be obtained, from the 438 proteins replicating in at least two out of the three injections. Using the absolute quantitation functionality of the software, 96% (480 ng) of the theoretical loading of 500 ng accounted for the 438 proteins.

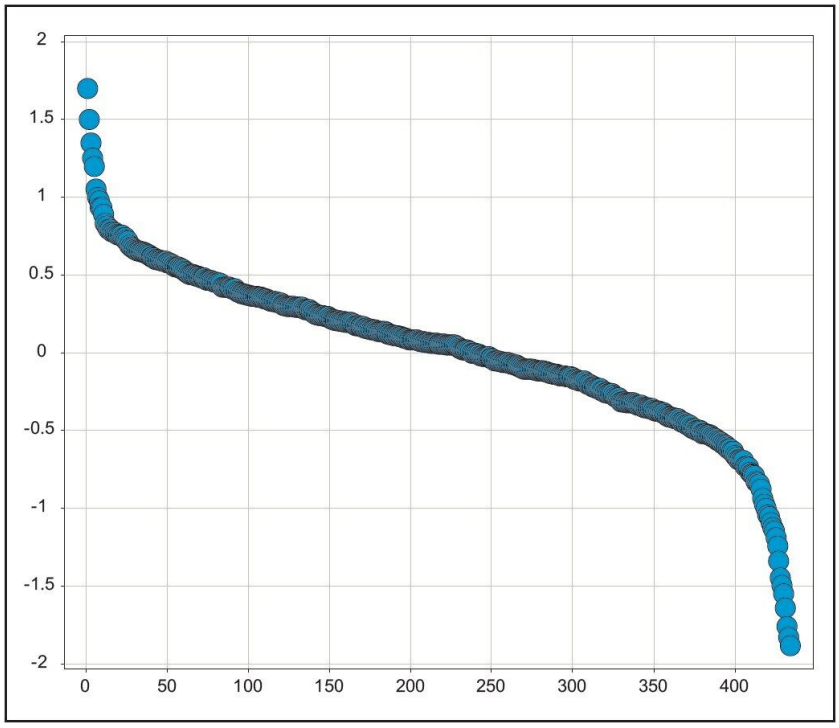


Figure 8. Absolute quantity for each one of the identified E. coli proteins detected on column. The amount is displayed in ng.

Conclusion

- The use of a parallel LC-MS^E data acquisition strategy provides an unbiased and highly reproducible approach to proteomic data acquisition, alleviating the issues seen with other, current LC-MS/MS approaches.
- The use of the LC-MS^E acquisition strategy in combination with a new databank searching strategy utilizing not just the MS information, but also physiochemical properties, results in a high degree of reproducibility and reliability in the results obtained.
- The very high degree of replication both at the peptide and protein level illustrates the ability of the strategy to mine deeper in to the data than other protein identification strategies.
- The statistical information generated by the LC-MS^E acquisition strategy allows the fine tuning of the empirical models, from tens of thousands of peptides on an injection-by-injection basis. This capability provides for very high selectivity, specificity, and sensitivity.

- The specificity of the scoring and validation process is remarkable and is illustrated by the level of replication of the identified *E. coli* proteins and the absence of false positive identifications in each of the six different bacteria proteomes.

References

1. Silva *et al.*, Mol. Cell. *Proteomics*. 5 (2006) 144–156.
2. Silva *et al. Anal. Chem*. 77 (2005) 2187–2200.
3. Krämer-Albers *et al., Proteomics Clin*. Appl. 1 (2007) 1446–1461.
4. Silva *et al. Mol. Cell. Proteomics*. 5 (2006) 589–607.
5. Hughes *et al. J. Proteome Res*. 5 (2006) 54–63.
6. Vissers *et al. Mol. Cell. Proteomics*. 5 (2007) 755–766.
7. Schwarz *et al., J. Sep. Sci*. 30 (2007) 2190–2197.
8. Levin *et al. J. Sep. Sci.* 30 (2007) 2198–2197.
9. Donoghue *et al. Proteomics* 6 (2008) in press.

Featured Products

ACQUITY UPLC M-Class System <https://www.waters.com/134776759>

720002631, May 2008