Waters™

アプリケーションノート

A Fully Automated Software Strategy for *de novo* Sequencing of Whole Q-Tof Electrospray LC-MS/MS Datasets

Iain Campuzano, Keith Richardson, Therese McKenna, Alistair Wallace, James I. Langridge

Waters Corporation

Abstract

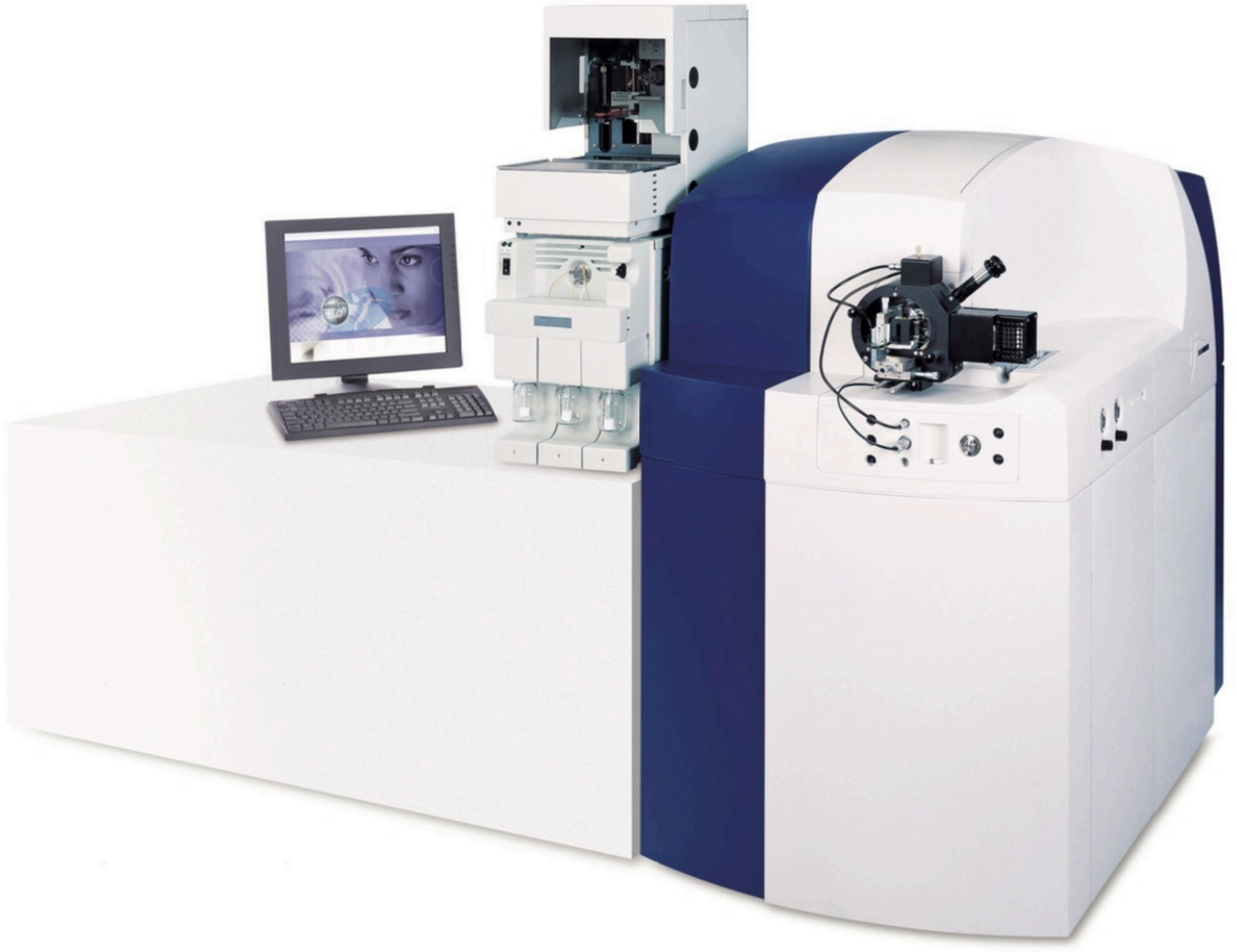
This application note describes the use of a Bayesian probability based *de novo* sequencing algorithm automatically applied to an entire LC-MS/MS dataset. The resultant sequences are compared to the results obtained from a databank search using the same dataset.

Introduction

Large numbers of proteins from sequenced organisms can be rapidly identified from single samples in an automated manner by LC-MS/MS analysis of the tryptic digests followed by databank searching. The enhanced sensitivity, resolution and mass measurement accuracy in data produced by the Waters Micromass Q-Tof Mass Spectrometer considerably aids the process of inferring *de novo* amino acid sequence.

The Maximum Entropy based MassSeq (ProteinLynx Global SERVER v2.0) *de novo* sequencing algorithm performs the following after raw spectra have been processed to remove isotope and charge multiplicity.

- A terminated Markov Chain Monte Carlo algorithm simulates the exploration of huge numbers of possible sequences by taking trial sequences and altering them in a pseudo-random manner to generate new trial sequences.
- The fragmentation process is modelled using Markov Chains. This allows the sum over all the possible fragmention patterns generated from the trial sequences to be calculated in linear time.
- The probability that each trial sequence accounts for the spectrum is calculated using Bayes' theorem.
- During the exploration, trial sequences are accepted or rejected according to this probability.



The Waters Micromass Q-Tof Mass Spectrometer.

Experimental

Data Acquisition

All data were acquired using a Waters Capillary HPLC (CapLC) System and a Q-Tof Ultima API, hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer (www.waters.com <http://www.waters.com>) fitted with a NanoLockSpray source. Data Directed Analysis (DDA) was carried out on 100 fmol of Yeast Enolase tryptic digest. Ions were selected for MS/MS analysis based on their intensity and charge state. Collision energies were chosen automatically based on the *m/z* and charge-state of the selected precursor ions. All data were acquired with an internal reference ion in order to provide mass measurement accuracies of less than 10 ppm in both the MS and MS/MS mode of acquisition.

Data Processing

ProteinLynx Global SERVER 2.0 was used to define a 'Workflow' template, comprising of a post-acquisition processing routine required to reduce the raw continuum MS/MS data set to a databank-searchable form, the databank to be searched and the associated query parameters (Figure 1).

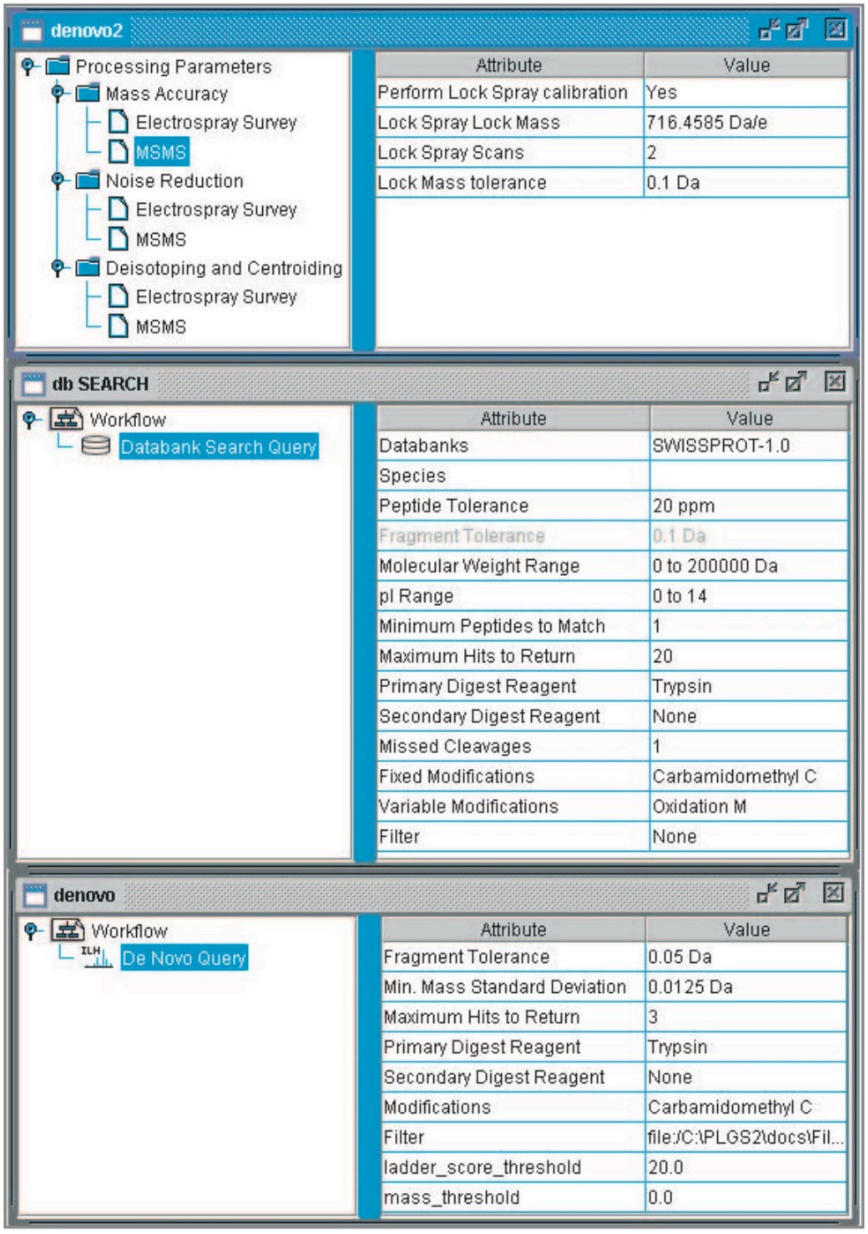


Figure 1. Template-driven data processing. Data processing parameter were set up in the Data Preparation tool. Databank searching and de novo sequencing parameters were specified in separate WorkFlow tempates.

A filter was applied to the dataset to discard those spectra containing insufficient information to represent a peptide. The remaining MS/MS spectra associated with each precursor ion were combined, transposed to a single charge state and reduced to a list of accurately mass measured peaks, using the MaxEnt 3 algorithm.

Precursor and product ions were automatically lockmass corrected against Glufibrinopeptide B and erythromycin respectively. All data was converted to an XML format, and searched against the SwissProt v.39 (86,865 entries, FASTA format) databank. An additional ‘Workflow’ was created, in which all spectra were *de novo* sequenced, with the results compared to the standard databank search. The 'Workflow' was automatically initiated upon completion of the LC-MS/MS acquisition, and the results displayed in an integrated Java interface (ProteinLynx Browser).

Results and Discussion

From 100 fmols of a Yeast Enolase tryptic digest, 59.6% sequence coverage was obtained from 25 unique peptides, identified by searching against SwissProt v.39. The RMS error of all precursor ions used for the identification of Yeast Enolase was 8.01 ppm. Table 1 shows a selection of peptides from the databank search (column 4) and the subsequent independent *de novo* sequencing (column 5) of the same peptides.

M/z	Charge state	Peptide mass	Database search sequence	De novo sequence
363.672	2	725.334	(R)SVYDSR(G)	SVYDSR
708.862	2	1415.714	(R)GNPTVEVELTTEK(G)	GNPTVEVELTTEK
362.228	2	722.444	(K)GVLHAVK(N)	GVLHAVK
643.851	2	1285.703	(K)NVNDVIAPAFVK(A)	NVNDVLAPAFVK
330.190	2	658.365	(K)ANIDVK(D)	ANLDVK
789.905	2	1577.794	(K)AVDDFLISLDGTANK(S)	AVDDFLLSLDGTANK
367.212	2	732.417	(K)NVPLYK(H)	NVPLYK
392.219	2	782.429	(K)HLADLSK(S)	HLADLSK
404.220	2	806.429	(K)TFAEALR(I)	TFAEALR
580.302	2	1158.603	(R)IGSEVYHNLK(S)	LGSEVYHNLK
644.854	2	1287.703	(K)VNQIGTLSESIK(A)	VNQLGTLSESLK
400.692	2	799.375	(K)YDLDFK(N)	YDLDFK
771.369	2	1540.722	-	(V)PSGASTGVHEALEMR
523.763	2	1045.519	-	(I)GSEVYHNLK
394.718	2	787.426	-	(F)MIAPTGAK
538.296	2	1074.592	-	(N)QIGTLSESIK

Table 1. Comparison between databank search and de novo sequencing. Column 4 shows a selection of peptide sequences identified by submitting data obtained by DDA to the SwissProt database. Column 5 shows the sequences obtained by de novo sequencing by MassSeq. Green - correctly assigned residues, Red - Incorrectly assigned, Yellow - I/L indistinguishable in low energy CID.

Two peptides *de novo* sequenced from the Yeast Enolase digest have been represented in Figure 2. Both peptides are doubly charged at *m/z* 392.22 and 580.31.

Figure 2a

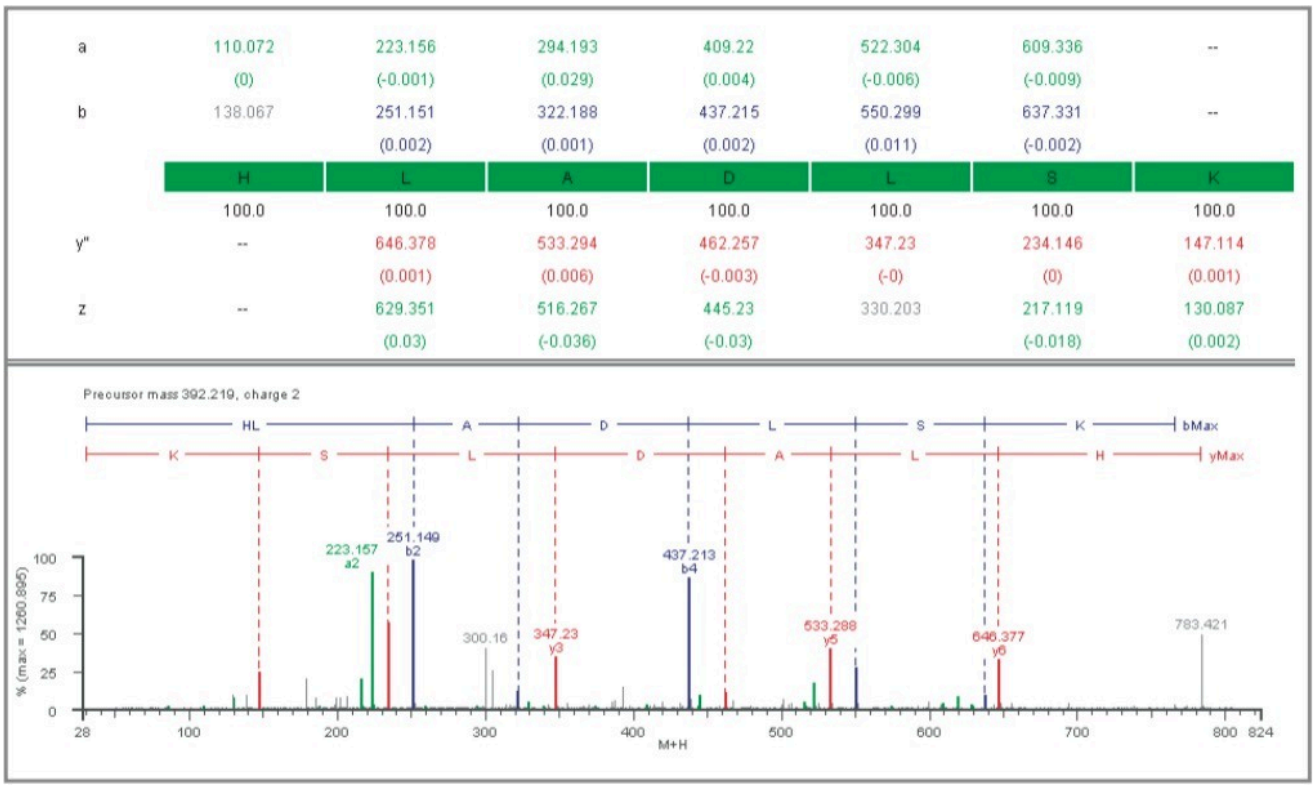

a 110.072 223.156 294.193 409.22 522.304 609.336 --
(0) (-0.001) (0.029) (0.004) (-0.006) (-0.009)
b 138.067 251.151 322.188 437.215 550.299 637.331 --
(0.002) (0.001) (0.002) (0.011) (-0.002)
H L A D L S K
100.0 100.0 100.0 100.0 100.0 100.0 100.0
y" -- 646.378 533.294 462.257 347.23 234.146 147.114
(0.001) (0.006) (-0.003) (-0) (0) (0.001)
z -- 629.351 516.267 445.23 330.203 217.119 130.087
(0.03) (-0.036) (-0.03) (-0.018) (0.002)
Precursor mass 392.219, charge 2
HL A D L S K bMax
K S L D A L H yMax
223.157 a2
251.149 b2
300.16
347.23 y3
437.213 b4
533.288 y5
646.377 y6
783.421
% (max = 1260.895)
M+H

Figure 2b

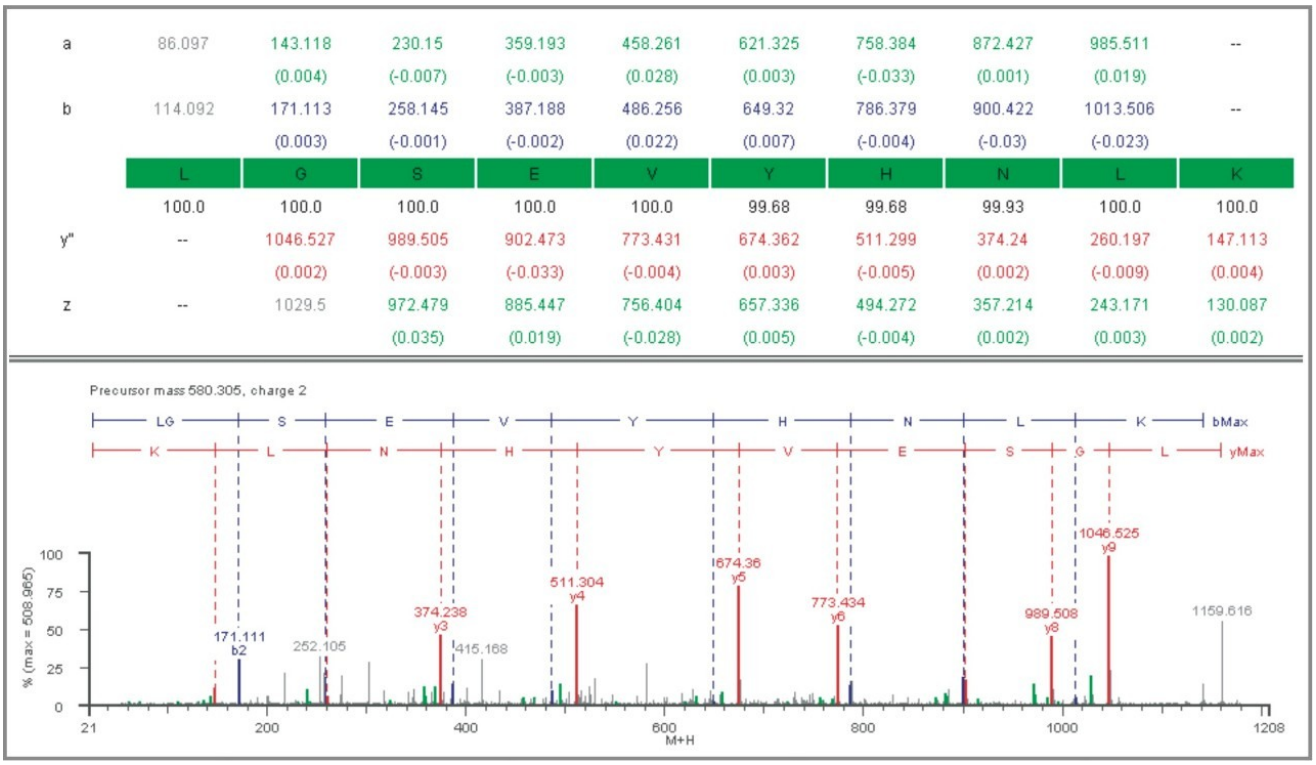


Figure 2. Full de novo sequence coverage of doubly charged peptides (a) 392.22 m/z and (b) 508.31 m/z from the LC-MS/MS experiment carried out on 100 fmol of injected Yeast Enolase. Also shown for each spectra are theoretical fragment masses for y", b, a and z ions, associated mass errors (Da) and MassSeq confidence level.

Upon *de novo* sequencing, the obtained sequence for ions *m/z* 392.22 and 580.31 were HLADLSK and LGSEVYHNLK respectively, which are consistent with the databank search results (Table 1). The four selected peptides of *m/z* 771.369, 523.763, 394.718 and 538.296 represent non-specific tryptic cleavages which were not identified by the standard databank search.

Conclusion

- The *de novo* sequencing algorithm generates high quality contiguous sequence information for long and short peptides, confirmed by the databank searching results.

- This approach can be used to identify those spectra unmatched by a databank search, due to non-tryptic cleavages or post-translational modification of the protein, thus increasing sequence coverage.
- The full automation of the *de novo* sequencing algorithm has been shown.
- The use of a NanoLockSpray source enables the mass accuracy obtained on the MS/MS fragments to be significantly enhanced (RMS error <10 ppm).

Featured Products

ProteinLynx Global SERVER (PLGS) <https://www.waters.com/513821>

720000724, August 2003