ApplicationNOTE

Waters

# 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 and James Langridge*
*Waters Corporation, Manchester, United Kingdom*

## Overview

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.

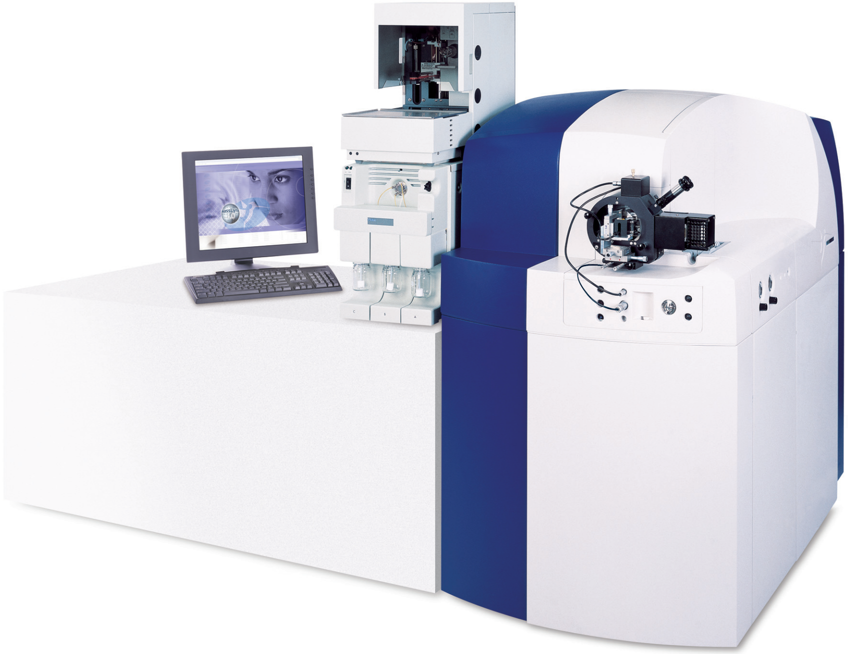
## Introduction

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 2.0) *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

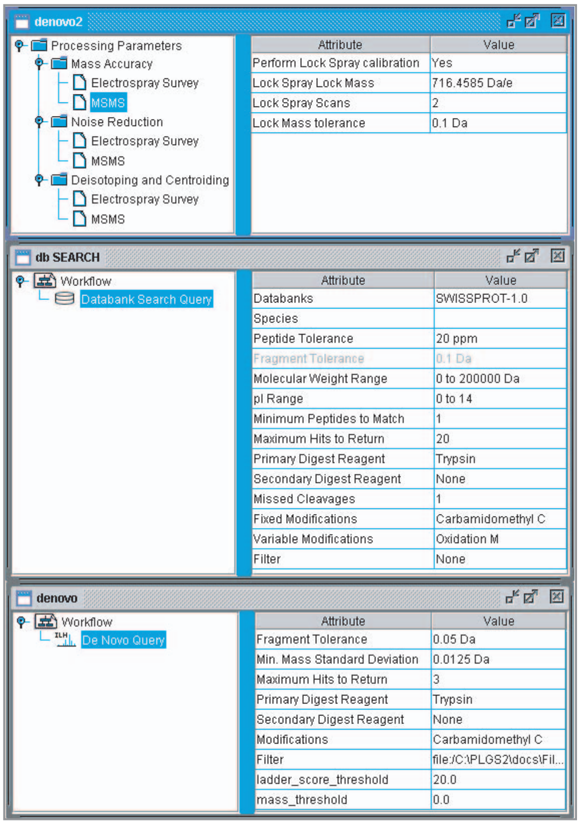
### *Data acquisition*

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) 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).

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 Glu-fibrinopeptide 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).



| Attribute | Value |
|---|---|
| Perform Lock Spray calibration | Yes |
| Lock Spray Lock Mass | 716.4585 Da/e |
| Lock Spray Scans | 2 |
| Lock Mass tolerance | 0.1 Da |

| Attribute | Value |
|---|---|
| Databanks | SWISSPROT-1.0 |
| Species | |
| Peptide Tolerance | 20 ppm |
| Fragment Tolerance | 0.1 Da |
| Molecular Weight Range | 0 to 200000 Da |
| pI Range | 0 to 14 |
| Minimum Peptides to Match | 1 |
| Maximum Hits to Return | 20 |
| Primary Digest Reagent | Trypsin |
| Secondary Digest Reagent | None |
| Missed Cleavages | 1 |
| Fixed Modifications | Carbamidomethyl C |
| Variable Modifications | Oxidation M |
| Filter | None |

| Attribute | Value |
|---|---|
| Fragment Tolerance | 0.05 Da |
| Min. Mass Standard Deviation | 0.0125 Da |
| Maximum Hits to Return | 3 |
| Primary Digest Reagent | Trypsin |
| Secondary Digest Reagent | None |
| Modifications | Carbamidomethyl C |
| Filter | file:/C:\PLGS2\docs\Fil... |
| ladder_score_threshold | 20.0 |
| mass_threshold | 0.0 |

*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.*

## Results

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 8.01ppm. 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.

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.

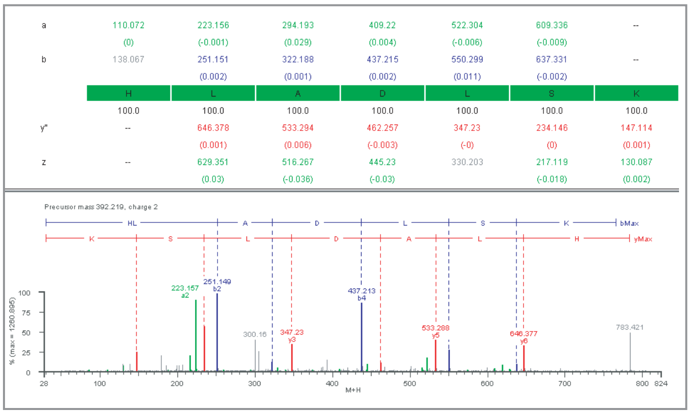


*Figure 2a*

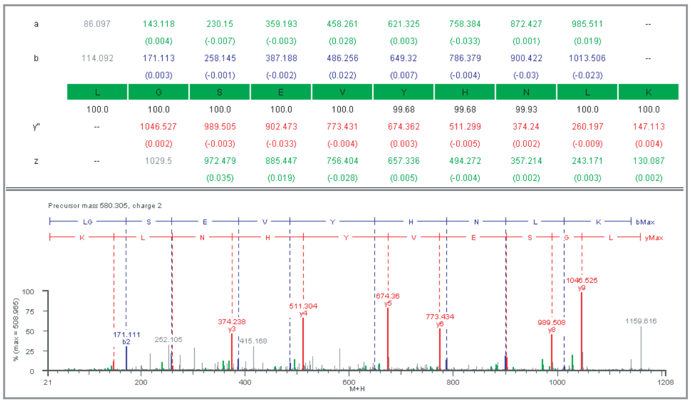


*Figure 2b*

*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.*

| 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.*

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 <10ppm).

## Sales Offices:

**AUSTRIA AND EXPORT** (CENTRAL EUROPE, CIS, MIDDLE EAST, INDIA AND INDIA SUBCONTINENT)
43 1 8771807

**AUSTRALIA** 61 2 9933 1777

**BELGIUM AND LUXEMBOURG**
32 02 726 1000

**BRAZIL** 55 11 5543 7788

**CANADA** 800 252 4752 X2205

**CIS** 7 095 931 9193

**CZECH REPUBLIC** 420 2 617 11384

**DENMARK** 45 46 59 8080

**FINLAND** 358 09 506 4140

**FRANCE** 33 1 3048 7200

**GERMANY** 49 6196 40 06 00

**HONG KONG** 852 29 64 1800

**HUNGARY** 36 1 350 5086

**INDIA** 91 80 837 1900

**IRELAND** 353 1 4481500

**ITALY** 39 02 274211

**JAPAN** 81 3 3471 7191

**KOREA** 82 2 3284 1300

**MEXICO** 52 55 5524 7636

**THE NETHERLANDS** 31 76 508 7200

**NORWAY** 47 6 384 6050

**PEOPLES REPUBLIC OF CHINA**
86 10 8451 8918

**POLAND** 48 22 833 4400

**PUERTO RICO** 1 787 747 8445

**SINGAPORE** 65 6278 7997

**SPAIN** 34 93 600 9300

**SWEDEN** 46 555 11 500

**SWITZERLAND** 41 62 889 2030

**TAIWAN** 886 2 2543 1898

**UK** 44 208 238 6100

**U.S.A. AND ALL OTHER COUNTRIES:**
WATERS CORPORATION
34 Maple St.
Milford, MA 01757 U.S.A.
T: 1 508 478 2000
F: 1 508 872 1990

WATERS CORPORATION
34 Maple St.
Milford, MA 01757 U.S.A.
T: 508 478 2000
F: 508 872 1990
www.waters.com/micromass

Produced in the United Kingdom

Waters
RIGHT ON TIME.

MICROMASS®
MS TECHNOLOGIES

LLOYD'S REGISTER QUALITY ASSURANCE
ISO9001

UKAS
QUALITY
MANAGEMENT
001