Waters

A FULLY AUTOMATED HIERARCHICAL SOFTWARE STRATEGY FOR *DE NOVO* SEQUENCING OF WHOLE Q-TOF ELECTROSPRAY LC-MS/MS DATASETS

James I. Langridge[1], A.Millar[1], P. Young[1], R. O'Malley[1], N. Swainston[1], J. Skilling[2], J. Hoyes[1], and K. Richardson[1]
[1] Waters Corporation, Floats Road Wythenshawe, Manchester.
[2] Maximum Entropy Data Consultants, Cambridge, UK.

Presented at ASMS 2002, Orlando, Florida, 3rd-6th June, 2002

Overview

- The aim of the following experiment was to demonstrate the utility of a *de novo* sequencing algorithm automatically applied to an entire LC-MS/MS dataset.
- A mixture of two proteins, present in equimolar amounts, was digested with trypsin and analysed by LC-MS/MS. The digest mixture was injected into, and separated by, a capillary liquid chromatography system prior to analysis with a Q-Tof *Ultima* API mass spectrometer.
- LC-MS/MS data were acquired using a nanoelectrospray source to provide exact mass measurement.
- All MS/MS data was acquired, processed and analysed in an automated manner.
- The MS/MS continuum data were filtered, combined, and transformed into singly charged, mass measured spectra.
- After conversion to XML, these spectra were *de novo* sequenced using a Bayesian probability based algorithm.
- These results were compared to the results of searching the same dataset against the Swiss-Prot database.

Introduction

Mass spectrometry has rapidly become the method of choice for the identification and characterisation of proteins. Large quantities of MS/MS spectra can be acquired and database-searched in an automated manner from LC-MS/MS experiments, allowing rapid identification of multiple proteins contained within a single sample.

A challenge remains, however, in that many of the acquired MS/MS spectra do not provide matches when searched against known protein sequences. The nature of database searching is such that only peptides whose molecular weight matches those within the databank will be identified. Consequently, many good quality spectra containing a single amino acid substitution remain unmatched. Furthermore, for organisms whose genomes are not completely sequenced electrospray ionisation tandem mass spectrometry (ESI-MS/MS) is an ideal technique since it can be used to provide high quality sequence data from individual peptides produced by the enzymatic digestion of protein. This high quality sequence information can either be used to probe existing protein or nucleotide databases using BLAST™ based homology searching or used for the generation of oligonucleotide primers (cDNA). The enhanced sensitivity, resolution and mass measurement accuracy in data produced by the Q-Tof *Ultima* API mass spectrometer considerably aids the process of inferring *de novo* amino acid sequence. Under favourable circumstances (*e.g.* tryptic peptides) it is possible to generate long unambiguous stretches of amino acid sequence.

In this work we describe a new fully automated program, that removes the need for manual intervention by automatically applying the *de novo* algorithm to an entire LC-MS/MS dataset. The resultant sequences are compared to the results of performing a database search using the same dataset. The entire process is controlled via a single Java interface.

Experimental

Data acquisition

All data were acquired using a Q-Tof *Ultima* API, hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer (www.micromass.co.uk).

The sample consisted of two proteins tryptically digested and mixed in equimolar amounts. 100 fmole aliquots of the samples were injected onto the LC-MS/MS set-up.

The analytical system used for the analysis consisted of a ten port valve (the stream select module) of a Micromass CapLC attached directly to the **Z** *SPRAY*™ source of the mass spectrometer. This was configured with a pre-concentration column in series with a nanoscale analytical column. The trapping column was packed with a C_{18} stationary phase (300 µm ID x 5 mm), the analytical column used was a 150mmx 75uM column (www.lcpackings.com) packed with PepMap C_{18} material.

This was set up to spray via a transfer capillary into the ESI source of the mass spectrometer through a nano-LC sprayer. Approximately 3000 volts were applied to the spray tip. A small amount of nebulising gas (approx. 5 psi) was also introduced around the spray tip to aid the electrospray process. A splitter gave a resultant flow through the analytical column of 200 nL/min with the pump programmed to deliver a flow of 2 µL/min.

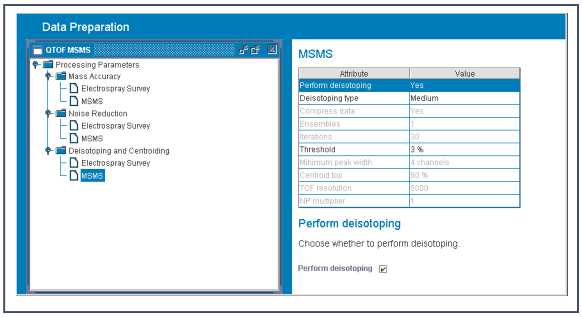
The mass spectrometer was operated in a data dependant acquisition (DDA) mode whereby following the interrogation of MS data, ions were selected for MS/MS analysis based on their intensity and charge state. Collision energies were chosen automatically based on the m/z and charge state of the selected precursor ions.

All data were acquired with an internal reference ion in order to provide mass measurement accuracies of less than 5 parts per million (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:

- the post-acquisition processing routines required to reduce the raw continuum MS/MS data set to a database-searchable form (**figure 1**);
- the *de novo* sequencing parameters (**figure 3**).
- the database to be searched and the associated query parameters (**figure 2**).



MSMS

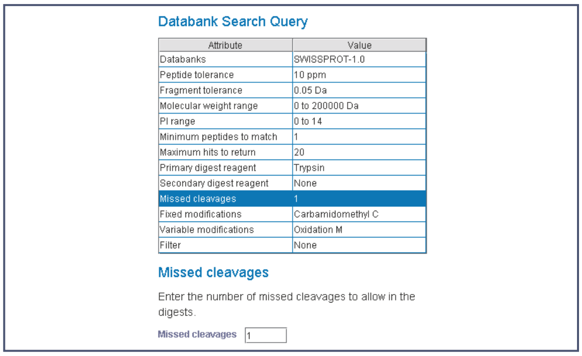
Attribute	Value
Perform deisotoping	Yes
Deisotoping type	Medium
Compress data	Yes
Ensembles	1
Iterations	30
Threshold	3 %
Minimum peak width	4 channels
Centroid top	80 %
TOF resolution	5000
NP multiplier	1

Perform deisotoping

Choose whether to perform deisotoping.

Perform deisotoping ☑

Figure 1. The post-processing parameter setup dialog



Databank Search Query

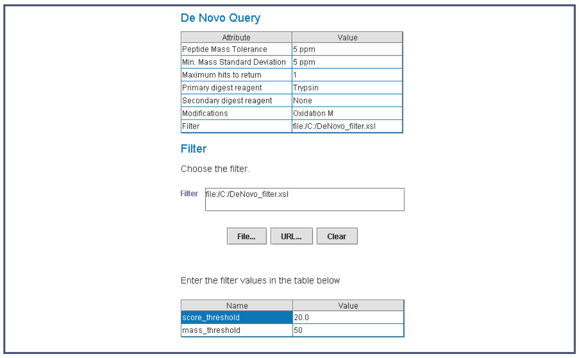
Attribute	Value
Databanks	SWISSPROT-1.0
Peptide tolerance	10 ppm
Fragment tolerance	0.05 Da
Molecular weight range	0 to 200000 Da
PI range	0 to 14
Minimum peptides to match	1
Maximum hits to return	20
Primary digest reagent	Trypsin
Secondary digest reagent	None
Missed cleavages	1
Fixed modifications	Carbamidomethyl C
Variable modifications	Oxidation M
Filter	None

Missed cleavages

Enter the number of missed cleavages to allow in the digests.

Missed cleavages 1

Figure 2. The database query setup dialog



De Novo Query

Attribute	Value
Peptide Mass Tolerance	5 ppm
Min. Mass Standard Deviation	5 ppm
Maximum hits to return	1
Primary digest reagent	Trypsin
Secondary digest reagent	None
Modifications	Oxidation M
Filter	file:/C:/DeNovo_filter.xsl

Filter

Choose the filter.

Filter: file:/C:/DeNovo_filter.xsl

Enter the filter values in the table below

Name	Value
score_threshold	20.0
mass_threshold	50

Figure 3. The de novo sequencing setup dialog

The 'Workflow' was automatically initiated on completion of the LC-MS/MS acquisition, and the results displayed in the ProteinLynx 2.0 Java interface.

The chromatogram from the LC ESI- MS/MS analysis is shown in **Figure 4** where it can clearly be seen that there are many peptides eluting from the column in the 25 minute period, between 10 to 35 minutes.

The data were processed and searched against the SwissProt v. 39 (86,865 entries, FASTA format) database according to the parameters defined in the WorkFlow template:

- 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 centroided using the MaxEnt 3 algorithm;
- after conversion to XML, the spectra were searched against the database and the returned hits validated using a post-search filter whereby a partial *de novo* sequence (amino acid tag) is generated for each matching spectrum and compared to the sequence returned from the database; a mismatch invalidates the hit.
- all spectra were *de novo* sequenced. The results are displayed in **Table 1**.

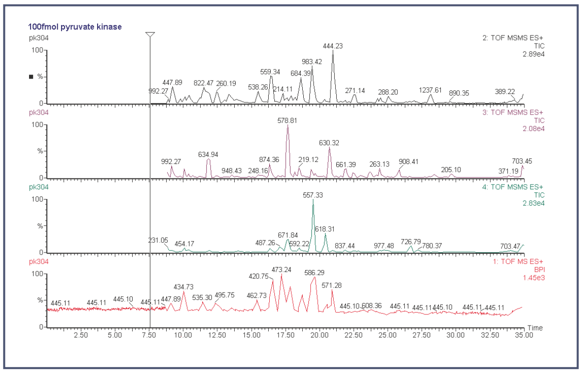


Figure 4. The de novo sequencing setup dialog

M/z	SwissProt result	mass delta ppm	*de novo* result	mass delta ppm	matching database entry
420.768	APIIAVTR	2.971	APLLAVTR	2.971	KPY1_RABIT
434.743	MQHLIAR	5.189	MQHLLAR	5.189	KPY1_RABIT
439.241	LFEELAR	-8.089	LFEVEQL	-8.089	KPY1_RABIT
448.241	ADDGRPFPQVIK	3.378	ADDQLLCELLPL	-10.85	ALFA_RABIT
469.234	AAQEEYVK	3.062	QEEYVRN	-23.784	ALFA_RABIT
470.745	ELSDIAHR	3.219	ELSDLAHR	3.219	ALFA_RABIT
473.259	VNLAMNVGK	9.486	VNLAMNVGK	9.486	KPY1_RABIT
495.758	GSGTAEVELK	2.591	GSGTAEVELK	2.591	KPY1_RABIT
503.289	KLFEELAR	3.078	KLFEELAR	3.078	KPY1_RABIT
510.263	GDYPLEAVR	-1.99	TAYPLEAVR	33.739	KPY1_RABIT
559.807	GSGTAEVELKK	-0.447	GSGTAEVELKK	-0.447	KPY1_RABIT
571.301	GDLGIEIPAEK	3.718	LGDGLELPAEK	3.718	KPY1_RABIT
577.292	SSGTSYPDVLK	-1.813	SSGTSYPDVLK	-1.813	TRY1_BOVIN
586.322	LDIDSAPITAR	-3.398	LDLDSAPLTAR	-3.398	KPY1_RABIT
588.997	KGVNLPGAAVDLPAVSEK	5.968	RELLPGAAVDLPAVESK	5.968	KPY1_RABIT
599.292	ITLDNAYMEK	5.324	LTLDNAYMKE	5.324	KPY1_RABIT
607.977	RFDEILEASDGIMVAR	0.843	MVWMLLEASDGLQLLC	-11.008	KPY1_RABIT
634.621	MNFSHGTHEYHAETIK	6.923	MNFSHGTFNMENMTLK	-6.545	KPY1_RABIT
666.853	GILAADESTGSIAK	2.11	LGLAADESTTELL	-6.323	ALFA_RABIT
708.027	QKGPDFLVTEVENGGFLGSK	7.879	KRDEEKVTEVENYLELQ	5.989	KPY1_RABIT
724.904	VYVDDGLISLQVK	-0.089	VYVDDGLLSLELL	-7.852	KPY1_RABIT
745.853	LQSIGTENTEENR	7.008	LQSLGTEEDEEKL	16.339	ALFA_RABIT
818.95	GVNLPGAAVDLPAVSEK	-0.764	VGNLPQAVDLAEGLLGA	-0.764	KPY1_RABIT
821.892	DPVQEAWAEDVDLR	-3.041	PDVQEAWAEDVDLR	-3.041	KPY1_RABIT
833.41	FDEILEASDGIMVAR	2.179	DFELLEAEELLECL	-11.318	KPY1_RABIT

Table 1. Comparison of the results obtained by searching the SwissProt databank and applying the de novo sequencing algorithm

m/z	*de novo* result	mass delta ppm	matching database entry
437.754	LVLTESGR	-0.448	KPY1_RABIT
457.272	PVAVALDTK	-0.396	KPY1_RABIT
521.751	AMGSVEASYK	-5.98	KPY1_FELCA
566.791	LAANSLRMQE	3.75	ALFA_RABIT
1053.035	YSLEYLVTEVENRVYDL	-8.557	KPY2_RABIT

Table 2. de novo sequencing results obtained for peptides not matched by searching the SwissProt database. blastp searching of the sequences produced the protein matches displayed

Results

Of the 30 ions searched against the database, 25 were matched to either Pyruvate Kinase or Fructose-bisphosphate Aldolase A. These assignments were validated by the post-search filter.

- For 11 ions, the *de novo* algorithm produced identical sequences to those returned by database searching; differing only by I or L (these residues being isobaric and therefore indistinguishable by low energy MS/MS). See example in **Figure 5**.
- In 14 cases, a high degree of homology was observed.
- For the 5 ions for which no database entry was found, the sequence produced by the *de novo* algorithm was searched using blastp (www.ncbi.nlm.nih.gov) against nrDB (www.ncbi.nlm.nih.gov), and each was matched to one of the two protein components of our mixture - Pyruvate Kinase or Fructose bisphosphate Aldolase A.
- blastp matches the sequence (**Figure 6**) from the first of the unmatched ions from **Table 2** to Pyruvate Kinase. This match was not made during the initial database search as the cleavage is non-tryptic:

    (L)IVLTESGR

- Similarly, the second of the unmatched ions (**Figure 6**) is also matched by blastp to Pyruvate Kinase. This match was not made during the initial database search as the cleavage is again non-tryptic:

    (R)PVAVA LDTK

- Similarly for **521.751**…

    (M)AMGSVEASYK

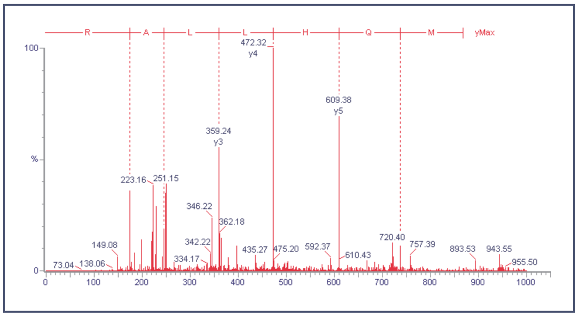


Figure 5. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 434.7, annotated against the peptide sequence *<u>MQHLLAR</u>*

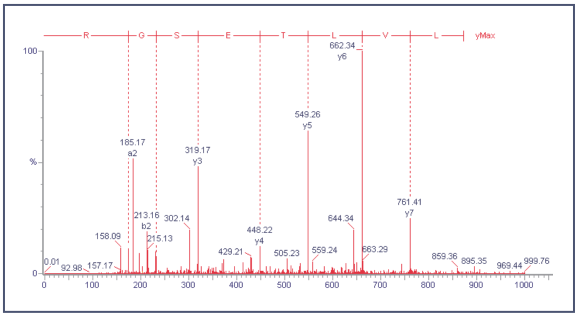


Figure 6. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 437.7, annotated against the peptide sequence *<u>IVLTESGR</u>*

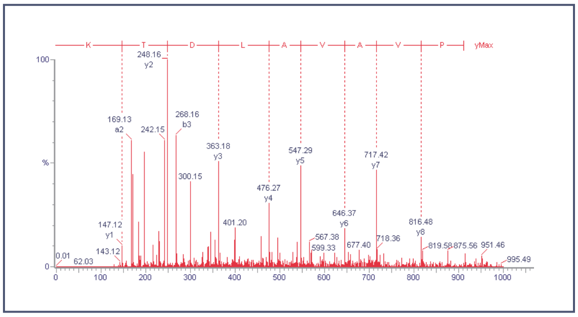


Figure 7. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 457.2, annotated against the peptide sequence *<u>PVAVALDTK</u>*

Conclusion

The full automation of a *de novo* sequencing algorithm has been shown.

Results from database searching and *de novo* sequencing have been compared and contrasted.

This approach can been used to identify those spectra initially unmatched by a database search, due to non-tryptic cleavages of the protein.

Future work will investigate the use of an automated BLAST routine in combination with the *de novo* sequencing algorithm.

PosterREPRINT

Author to whom all correspondence should be addressed:
Alan Millar
Waters Corporation
(Micromass UK Limited)
Floats Road, Wythenshawe
Manchester, M23 9LZ
Tel: + 44 (0) 161 946 2400
Fax: + 44 (0) 161 946 2480
e-mail: alan.millar@micromass.co.uk

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

Made in the United Kingdom

Waters
RIGHT ON TIME.

MICROMASS®
MS TECHNOLOGIES

For research use only. Not for use in diagnostic procedures.

LLOYD'S REGISTER QUALITY ASSURANCE ISO9001
UKAS QUALITY MANAGEMENT 001
TickIT
Certificate No: 951387