PosterReprint

Waters

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

J. Langridge[1], A. Wallace[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] Micromass UK Ltd, Floats Road Wythenshawe, Manchester. [2] Maximum Entropy Data Consultants, Cambridge, UK.

Presented at EPS 2002, Sorrento, Italy, 1st - 7th September 2002

OVERVIEW

To demonstrate the utility of a *de novo* sequencing algorithm as automatically applied to an LC-MS/MS dataset.

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 that match exactly 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 interpret 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 assessing the results of a database search as performed on an entire LC-MS/MS data file, and applying the *de novo* algorithm to any unmatched spectra. 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 C18 stationary phase (300µm ID x 5mm), the analytical column used was a 150mm x 75µm column (www.lcpackings.com) packed with PepMap C18 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 (approx. 5-psi) of nebulising gas was also introduced around the spray tip to aid the electrospray process.

Waters

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 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 database to be searched and the associated query parameters (**Figure 2**);
- the filtering process whereby a subset of the processed and searched MS/MS spectra were passed to the *de novo* algorithm.
- the *de novo* sequencing parameters (**Figure 3**).

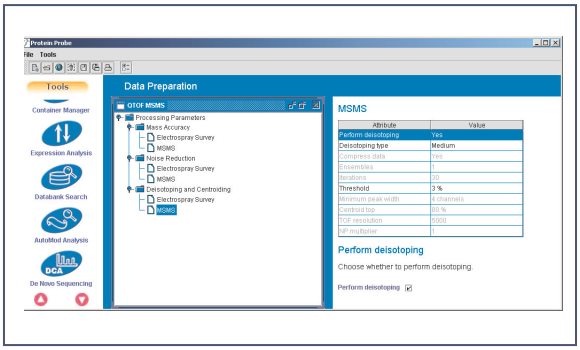

Figure 1. The post-processing parameter setup dialog

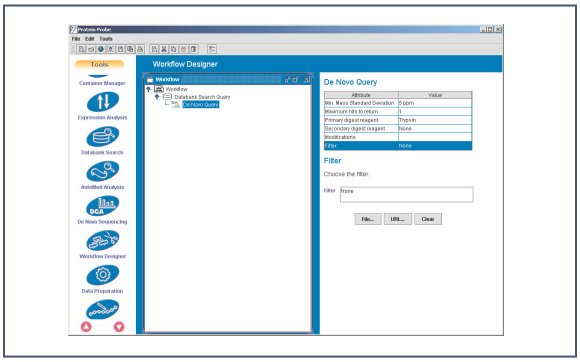

Figure 2. The database query setup dialog

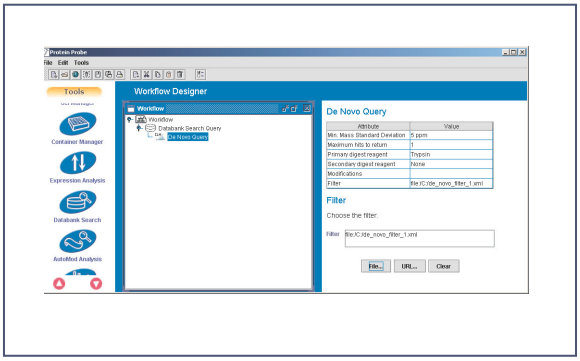

Figure 3. The de novo sequencing setup dialog

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

Waters

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.

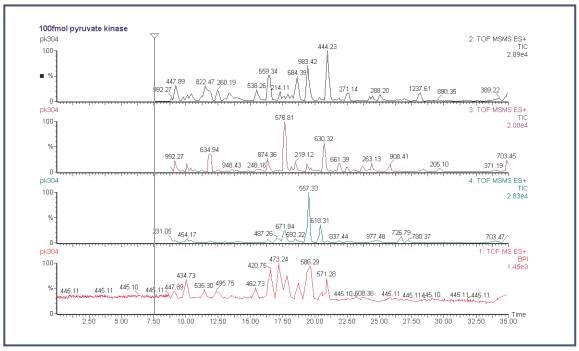

Figure 4. LC ESI chromatograms

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 and then transformed to a single charged, mass measured spectrum using the MaxEnt 3 algorithm;
- after conversion to XML, the spectra were database searched against the database and returned hits validated using a post-search filter, in which a number of partial amino acid sequences (tags) were generated for each spectrum. Returned sequences not containing one or more of these tags were rejected.
- all spectra were *de novo* sequenced. The results are displayed in **Table 1**.

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	KGVNLPGAAVDL PAVSEK	5.968	RELLPGAAVDLP AVESK	5.968	KPY1_RABIT
599.292	ITLDNAYMEK	5.324	LTLDNAYMKE	5.324	KPY1_RABIT
607.977	RFDEILEASDGIM VAR	0.843	MVWMLLEASD GLQLLC	-11.008	KPY1_RABIT
634.621	MNFSHGTHEYHA ETIK	6.923	MNFSHGTFNME NMTLK	-6.545	KPY1_RABIT
666.853	GILAADESTGSIAK	2.11	LGLAADESTTEL L	-6.323	ALFA_RABIT
708.027	QKGPDFLVTEVEN GGFLGSK	7.879	KRDEEKVTEVE NYLELQ	5.989	KPY1_RABIT
724.904	VYVDDGLISLQVK	-0.089	VYVDDGLLSLEL L	-7.852	KPY1_RABIT
745.853	LQSIGTENTEENR	7.008	LQSLGTEEDEEK	16.339	ALFA_RABIT
818.95	GVNLPGAAVDLP AVSEK	-0.764	VGNLPQAVDLA EGLLGA	-0.764	KPY1_RABIT
821.892	DPVQEAWAEDVD LR	-3.041	PDVQEAWAEDV DLR	-3.041	KPY1_RABIT
833.41	FDEILEASDGIMV AR	2.179	DFELLEAEELLE CL	-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	YSLEYLVTEVENR VYDL	-8.557	KPY2_RABIT

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

Waters

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 BLAST-P (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 and each was matched to one of the two protein components of the mixture: pyruvate kinase or fructose-biphosphate aldolase A.

- BLAST-P 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 7**) is also matched by BLAST-P 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

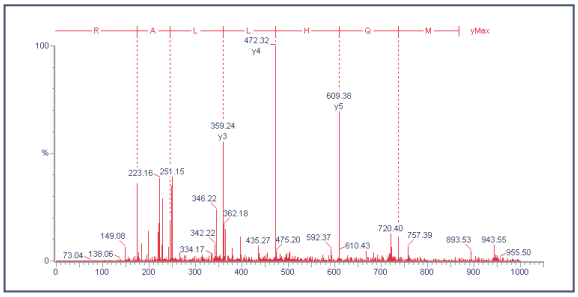

Figure 5. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 434.7, annotated against the peptide sequence MQHLLAR

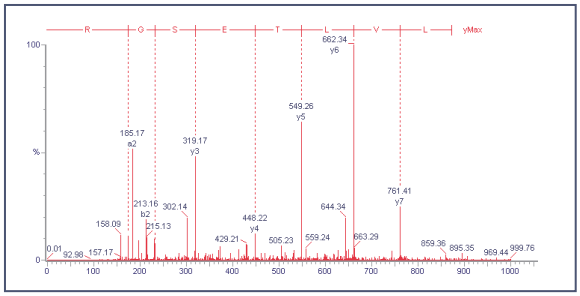

Figure 6. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 437.7, annotated against the peptide sequence IVL TESGR

Waters

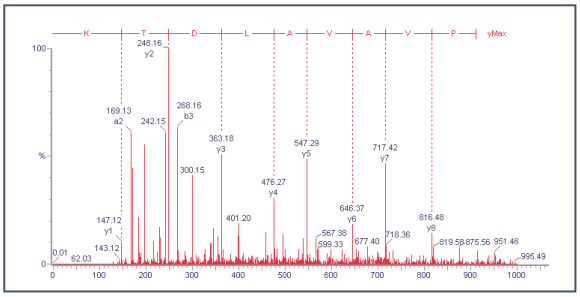

Figure 7. Electrospray MS/MS spectrum for the doubly charged precursor ion m/z 457.2, annotated against the peptide sequence PVAVA LDTK

CONCLUSION

The full automation of a *de novo* sequencing algorithm subsequent to database searching has been shown.

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

This approach has 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

MS Sales Offices:

BELGIUM 02-2534550

CANADA 514 694-1200

DENMARK 4657 4101

EUROPE +31 (0) 36-540 6000

FINLAND 02 284 56 11

FRANCE 0800-907016

GERMANY 0800-1817249

ITALY 02 2159 1415

NETHERLANDS 036-540 6160

NORDIC +46 (0) 8 555 115 10

SPAIN 93 440 71 30

SWEDEN 08 555 115 10

SWITZERLAND (FRENCH) 0800-558334

SWITZERLAND (GERMAN) 0800-556190

UK 0161 435 4125

USA 978 524-8200

To whom all correspondence should be addressed:
Tel: + 44 (0) 161 435 4100
Fax: + 44 (0) 161 435 4444
e-mail: alistair.wallace@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

Waters
RIGHT ON TIME.

MICROMASS®
MS TECHNOLOGIES

Renewable Fuels Association, www.ethanolrfa.org
California Energy Commission Report, P600-01-017, August 2001.

Certificate No: 951387