ASMS 2004

A Novel Approach to Protein Identification: A Direct Comparison to Traditional Mass Spectrometric Techniques

Thérèse McKenna, Iain Campuzano, Mark Ritchie, Phillip Young, Keith Richardson, Richard Denny, Kieran Neeson, Christopher Jones and James Langridge
Waters Corporation, Manchester, UK

Introduction

A common goal of many analytical studies in the life sciences is the qualitative identification of the constituent proteins present in biological samples. Recent advances in both HPLC and mass spectrometry have allowed the analysis of complex protein mixtures, which have not been separated on a 2D gel. These experiments normally involve tryptic digestion, followed by separation of the complex peptide mixture by microcapillary liquid chromatography, connected to an electrospray mass spectrometer capable of data dependant switching between the MS and MS/MS modes. In these experiments the mass spectrometer will perform classic MS/MS product ion acquisition on one precursor ion at a time with selection of the precursor in a serial manner using the first mass analyser. Protein identification is then achieved via database searching of the associated ESI-MS/MS data. Using this approach it has been demonstrated that hundreds of MS/MS spectra can be acquired in a fully automated fashion, resulting in the identification of significant numbers of proteins from a single LC/MS/MS experiment.

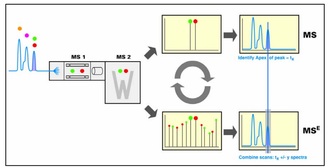


Figure 1. *Schematic of the novel LC/MS experiment on the electrospray Q-Tof mass spectrometer. Throughout the data acquisition the quadrupole functions in RF only mode.*

*Here we present a novel approach where the Q-Tof mass spectrometer is programmed to cycle between low and elevated collision energy acquisitions, during the course of an LC/MS run (**Figure 1**). The resolution and mass measurement accuracy of the oa-TOF mass analyser is key to this approach as it provides exact mass data. **At no point during the experiment are precursor ions selected with the quadrupole, as in a traditional product ion MS/MS acquisition.** The low energy MS dataset provides the exact mass of all detectable peptide molecular ions, whilst parallel fragmentation of all ions that are present within the hexapole gas cell at any particular time point is obtained in the elevated energy data. This results in the fragmentation of every detectable peptide precursor present in the low energy MS acquisition. From this data, amino acid sequence information can be obtained.*

An added benefit of this parallel analysis is that the chromatographic integrity of the peptides is maintained and the measurement of peptide ion intensity is significantly more reproducible from one experiment to the next, as a greater sampling rate result in more data points across the chromatographic peak. This allows the comparison of relative expression levels of identical proteins between two or more samples, without the need to use stable isotope labelling.

Experimental

We compare and contrast traditional mass spectrometric approaches with this novel method for the analysis of a simple protein digest mixture. In addition we demonstrate how protein identification and relative quantification is achieved without the use of traditional LC/MS/MS acquisitions. To this end we utilise proprietary software, developed to process this data in a quantitative and qualitative manner.

All data was acquired using the Waters® Protein Expression System under the following conditions:

Chromatography

- *All HPLC was performed on a Waters CapLC® System with a Stream Select Module providing direct flow at a rate of 9µL/min, split to 300nL/min.*
- *Peptides were initially trapped on a Waters Symmetry300™ OPTI-PAK™ Trapping Column and subsequently separated using a 15 cm x 75 µm, 3.5 µm NanoEase™ Atlantis™ dC$_{18}$ Column.*
- *An acetonitrile gradient from 5% to 40% over 90 minutes eluted the peptides from the analytical column.*

Mass Spectrometry

- *LC/MS data was acquired in continuum mode over the m/z range 50-1990 using a Waters Micromass® Q-Tof Ultima API, operated at a mass resolving power of 17500 (FWHM) in W-optics™ mode.*
- *Low (10eV) and elevated (23-33eV) collision energy scans were acquired in an alternate fashion for 1.5 seconds each, with an inter-scan delay of 0.15 seconds.*
- *NanoLockSpray™ was used to enable exact mass acquisitions. [Glu]^1-Fibrinopeptide B (100 fmol/µl) and Erythromycin (50 fmol/µl) were introduced through the lock mass channel and a single-point was used for lock mass correction of the acquired data.*
- *Data Directed Analysis (DDA™) employed automatic MS/MS switching on the four most intense multiply charged precursor ions in the m/z range 350-1600 over the m/z range 50-1990.*

Expression Bioinformatics

*For a given data file, ions are detected using an algorithm that employs a maximum likelihood technique to determine the exact mass, intensity, retention time and precision estimates thereof for each eluting peptide species. This output of exact mass retention time pairs [EMRTs] can be used to match and compare expression levels from different sample injections. Further processing allows the exact mass values for peptide ions taken from the low energy data and their corresponding fragment ions, determined from the elevated energy data, to be searched against a databank using a proprietary peptide fragmentation model[1]. The result is that each protein in the databank is given a probability score and confidence value. These results are displayed in an interactive browser and list the matched protein sequences ranked by their probability. This display also allows visualisation of the associated peptide sequences, which match to each protein and displays the relevant fragment ion data. The expression ratio of identified proteins having a significant probability under different conditions can then be determined. The probabilistic technique also provides a measure for the uncertainty of the ratios. These proprietary methods described above are referred to as **Waters® Protein Expression System Informatics** and are optionally incorporated into ProteinLynx Global Server 2.2. DDA data was processed and peptides identified using ProteinLynx Global Server 2.2, SwissProt v 40 was used for all searches.*

Sample Information

*Tryptic digests of Enolase (Yeast), Alcohol Dehydrogenase (Yeast), Serum Albumin (Bovine), Glycogen Phosphorylase B (Rabbit) and Hemoglobin (Bovine) (**Waters Protein Expression System Standards**) were prepared and analysed as mixtures (**Table 1**). These were also spiked into an E.coli cytosolic fraction tryptic digest.*

	Sample 1a	Sample 1b	Sample 2	Sample 3
	fmol injected			
Glycogen Phosphorylase (rabbit)	100	1000	100	1000
Serum Albumin (bovine)	100	10,000	250	100
Enolase (yeast)	100	10	500	10
Alcohol Dehydrogenase (yeast)	100	100	250	250
Hemoglobin (bovine)	-	-	1000	10,000

Table 1. *Standard protein matrix, used to assess the Protein Expression System via DDA qualitative comparisons and quantitative analysis.*

Results

*Samples 1a and 1b (see **Table 1**), were analysed by both LC/MS/MS (using Data Directed Analysis) and the novel LC/MS based approach to compare the number of proteins identified and the protein sequence coverage of the two techniques. **Figure 2** shows typical low and elevated energy chromatograms, generated by this approach. Inset, are the low energy and associated elevated energy mass spectra obtained at 37.3 minutes.*

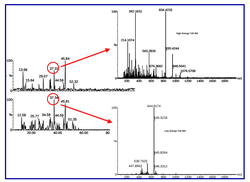


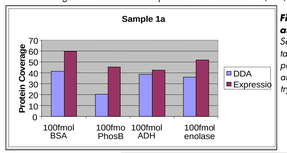
Figure 2. *Typical chromatograms generated by the Waters Protein Expression System approach with (inset) low energy and the associated elevated energy mass spectra at 37.3 minutes.*

*The quality of fragment ion spectra from both methods was compared to assess both reliability of the peptide identification and the sequence coverage of the protein. **Figures 3a and 3b** show fragment ion data obtained from an alcohol dehydrogenase tryptic peptide. The fragment ion mass spectrum is more complex in the expression analysis data, however, comparable fragmentation of the peptide is achieved when compared to traditional DDA. The correct assignment of these fragment ions is made possible by the high resolution and excellent mass measurement accuracy of the instrument, despite the complexity of the fragment ion spectrum.*

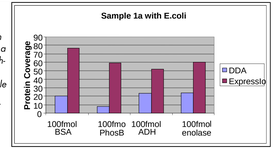
Figure 3a. *Parallel fragmentation of the doubly charged ion from alcohol dehydrogenase at m/z 724.4, eluting at 54.37 mins. Shown are the annotated sequence ions and fragment ion mass errors.*

Figure 3b. *Traditional MS/MS fragmentation produced by the DDA approach of the doubly charged ion at m/z 724.4. Shown are the annotated sequence ions and fragment ion mass errors.*

*Further analysis of the data shows that both approaches have the ability to characterise the equimolar protein sample, present as a simple mixture or in the complex E.coli fraction. However, the novel Protein Expression System approach demonstrates a significant increase in sequence coverage of the proteins, compared to the traditional LC/MS/MS DDA analysis. (**Figure 4a**). As the complexity of the sample is increased by the addition of E.coli digested proteins (**Figure 4b**), coverage of the standard proteins by DDA is approximately halved due to the serial nature of precursor ion selection and the decreased likelihood of selecting peptide ions from the standard proteins. Expression profiling, however, is not limited by having to select precursor ions with MS1 in a serial manner, as all the ions are fragmented simultaneously. Consequently no decrease in protein coverage is observed when the sample complexity is increased. In general, the Protein Expression System analysis provides better coverage on the four standard proteins than traditional LC/MS/MS.*



Figures 4a. (left) and 4b. (below) *Sequence coverage obtained for the standard proteins in the absence and presence of E.coli tryptic lysate.*



*When the proteins are present at different concentrations, ranging from 10 to 10,000 fmoles, in a simple mixture, both methods identify all the proteins present in the sample (**Figure 5a**) but again protein coverage by LC/MS/MS is significantly reduced for each of the spiked digest standards.*

*As found previously when the complexity is increased by the addition of E.coli tryptic peptides, the coverage of the spiked standard proteins is further reduced when analysed by LC/MS/MS. (**Figure 5b**) This effect is not observed with the Protein Expression System analysis and the dynamic range present in the sample does not adversely affect the coverage of the proteins. It is evident that increasing the sample complexity heavily affects the coverage of those proteins at lower concentration when the sample is analysed by LC/MS/MS (**Figure 5b**). Peptides from Enolase (10 fmols injected) are not discernible by either method, perhaps due to the increased complexity introduced by the E.coli tryptic digest.*

Quantification

*Quantification of the standard proteins in the presence of the E.coli tryptic peptides was performed by comparing two of the standard samples, Samples 2 and 3 (shown in **Table 2**). The first step in this quantification process is the identification of the constituent proteins. Data was processed and searched against the Swiss-Prot databank and resulted in the identification of all five of the protein standards with high confidence and good sequence coverage. In addition over 100 E.coli proteins were identified. **Table 2** shows the confidence, coverage and root mean square (RMS) mass errors calculated from all of the peptides for the five standard proteins obtained from this qualitative identification.*

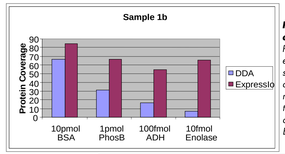
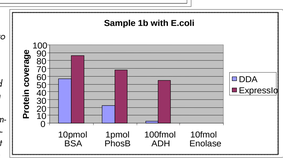


Figure 5a. (left) and 5b. (below) *Protein sequence coverage obtained from a sample of the standard proteins at a range of concentrations in the absence and presence of an E.coli tryptic digest.*

*To enable relative quantification the two datsets were then normalised by the software, using the confidently identified tryptic peptides from one of the standard proteins (in this example Alcohol dehydrogenase). Subsequent quantification of the remaining protein standards and their up and down regulation is displayed in the browser screenshot in **Figure 6**. The measured and expected change in concentration, [analogous with expression], for these proteins is shown in **Table 3**. The concentrations of Glycogen Phoshorylase b and Hemoglobin are both increased ten fold and this is evident from results shown by Protein Expression System profiling; where Glycogen Phosphorylase b is up regulated by 10.2 times, and the alpha and beta chains of hemoglobin raised by over 9 fold. Serum albumin was observed as being down regulated two fold, as expected. Enolase, was not observed in one of the samples and could not be quantified but does appear as a "unique" protein for the sample.*



Protein	RMS error	% Probability	# of Peptides	% Coverage
Serum Albumin P02769	2.93 ppm	100	43	63.4
Enolase P00924	5.08 ppm	100	25	57.8
Glycogen Phosphorylase B P00489	3.24 ppm	100	41	38.7
Alcohol Dehydrogenase P00330	3.40 ppm	100	15	42.4
Hemoglobin Beta chain P02070	2.69 ppm	100	12	69.7
Hemoglobin Alpha chain P01966	1.26 ppm	100	7	51.8

Table 2. *Standard proteins identified from Sample 1, using informatics analysis. The accompanying mass errors are the root mean square (RMS) errors of the errors of all the identified peptides from that protein.*

*Proteins identified and quantified from these two experiments were plotted using Spotfire (**Figure 7**). The vertical axis is the change in area of the same protein between the two samples [analogous to the up or down regulation of a given protein). The horizontal axis is a plot of the P number (a product of the standard deviation of the change in ion intensities) associated with a specific protein identification. The points that are parallel to the floor, along the central axis, represent the E.coli proteins and ADH (the internal standard), which have not changed in their expression ratio. The up-regulated proteins are shown on the left hand wall separated by their p value at 1 (indicating significant and reproducible up regulation). Down regulated proteins are seen on the right hand wall, separated by their p value being at or close to zero, indicating significant and reproducible down regulation.*

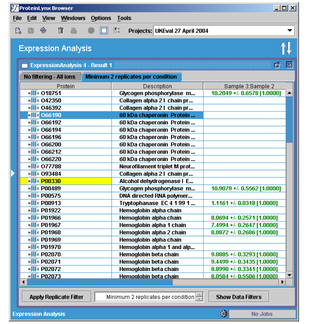


Figure 6. *Quantification results for standard protein digests in the presence of the E.coli tryptic lysate digest.*

Protein	Expression Ratios Sample 3 : Sample 2	
	Theoretical	Measured
Serum Albumin P02769	0.4	0.5281 ± 0.0219
Enolase P00924	-	-
Glycogen Phosphorylase B P00489	10	10.9079 ± 0.5562
Alcohol Dehydrogenase P00330	Internal Standard	Internal Standard
Hemoglobin Beta chain P02070	10	9.0885 ± 0.3293
Hemoglobin Alpha chain P01966	10	8.0694 ± .02571

Table 3. *Informatics analysis results. Proteins and expression levels obtained directly from the informatics browser compared with the theoretical levels.*

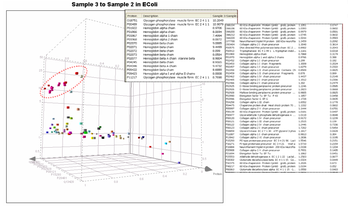


Figure 7. *Spotfire plot for quantified comparison of Samples 3 and 2 in the presence of the E.coli tryptic lysate digest.*

Discussion

The Waters Protein Expression System approach is a paradigm shift in LC/MS analysis of protein digests. It is not restricted to obtaining fragment ion data on individual precursor ions, instead it fragments all peptide precursor ions simultaneously.

The data indicate that the Protein Expression System approach was able to provide better sequence coverage for lower concentration proteins than conventional DDA MS/MS analysis.

This strategy also provides the ability to quantitively measure the relative level of proteins contained in two or more complex protein digest samples. This quantification is possible without chemical derivatisation of the peptides. Moreover both qualitative and quantitative information is obtained in a single exact mass LC/MS experiment.

Summary

- *The Waters Protein Expression System combines excellent chromatographic reproducibility, the exact mass measurement of precursor ions and associated fragment ions with proprietary bioinformatics to generate quality protein identifications.*
- *This approach provides increased protein coverage compared to a traditional LC/MS/MS experiments.*
- *Increasing the sample complexity has little or no effect on protein sequence coverage when using the expression system.*
- *Quantitive measurements of the relative level of proteins contained in two or more complex protein digest samples can be made.*
- *This analysis can be performed without the need for derivatisation of a sample.*

Reference

1. *John Skilling, Richard Denny, Keith Richardson, Phillip Young, Therese McKenna, Iain Campuzano and Mark Ritchie. Probseq - a fragmentation model for interpretation of electrospray tandem mass spectrometry data. Comp Funct Genom 2004; 5: 61-68.*

720000899EN May 2004