Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

Journal of Experimental Botany, Vol. 57, No. 7, pp. 1509–1514, 2006
Plant Proteomics Special Issue
doi:10.1093/jxb/erj139 Advance Access publication 21 March, 2006

Journal of Experimental Botany

Reproducibility of LC-MS-based protein identification

Matthias Berg[1,*], Axel Parbel[1], Harald Pettersen[2], David Fenyö[2] and Lennart Björkesten[2]

[1] *GE Healthcare, Oskar-Schlemmer-Strasse II, D-80807 München, Germany*

[2] *GE Healthcare, Björkgatan 30, Uppsala, Sweden*

Received 8 July 2005; Accepted 26 January 2006

Abstract

Traditional analysis of liquid chromatography-mass spectrometry (LC-MS) data, typically performed by reviewing chromatograms and the corresponding mass spectra, is both time-consuming and difficult. Detailed data analysis is therefore often omitted in proteomics applications. When analysing multiple proteomics samples, it is usually only the final list of identified proteins that is reviewed. This may lead to unnecessarily complex or even contradictory results because the content of the list of identified proteins depends heavily on the conditions for triggering the collection of tandem mass spectra. Small changes in the signal intensity of a peptide in different LC-MS experiments can lead to the collection of a tandem mass spectrum in one experiment but not in another. Also, the quality of the tandem mass spectrometry experiments can vary, leading to successful identification in some cases but not in others. Using a novel image analysis approach, it is possible to achieve repeat analysis with a very high reproducibility by matching peptides across different LC-MS experiments using the retention time and parent mass over charge (*m/z*). It is also easy to confirm the final result visually. This approach has been investigated by using tryptic digests of integral membrane proteins from organelle-enriched fractions from *Arabidopsis thaliana* and it has been demonstrated that very highly reproducible, consistent, and reliable LC-MS data interpretation can be made.

Key words: DeCyder™ MS, differential expression analysis, LC-MS, reproducibility, reversed phase chromatography, nano LC, tandem mass spectrometry.

Introduction

Proteomics has the potential to make a major contribution in the quest to cure human disease by comparing the protein levels in healthy and diseased samples. This also includes the analysis of samples representing different stages of disease, and under differing biological conditions to understand more clearly the role that proteins play and to identify potential biomarkers. Mass spectrometry–based proteomics has the capability to identify hundreds of proteins in a single experiment, and has become an important analytical technology in modern biological and medical research (Aebersold and Mann, 2003).

Proteomics samples, which are often complex mixtures of proteins, are usually digested with an endoprotease such as trypsin before mass spectrometry analysis. In a classical liquid chromatography-mass spectrometry (LC-MS) experiment, the resulting peptides are then separated by reversed-phase micro- or nano-capillary chromatography. Peptides eluting from the LC column are usually ionized by electrospray and then introduced into the mass spectrometer. Peptide masses and intensities are measured with the mass spectrometer and based on the signal intensity peptides are selected for fragmentation to obtain information on their sequence. Tandem mass spectra are acquired and searched against sequence collections to identify the corresponding peptides and proteins (Fenyo, 2000).

The traditional way to visualize LC-MS data for data quality assessment and confirmation of results is to use total ion or base ion chromatograms together with single or averaged mass spectra of all peptides eluting at a certain time, and tandem mass spectra of single peptides. Such visualizations provide detailed insight into a specific performance characteristic of an LC-MS experiment, such as the quality of fragment spectra, mass resolution, or chromatographic peak resolution. However, they are not very intuitive, and the information on *m/z* and retention time correlation is not easily accessible. By contrast, with a two-dimensional visualization of LC-MS data, it is easy to find indicators for problems like non-covalent adduct formation or sample contamination (e.g. PEG) commonly encountered in LC-MS

* To whom correspondence should be addressed. E-mail: matthias.berg@ge.com

analyses (Schulz-Knappe *et al*., 2001; Heine *et al*., 2002; Palmblad *et al*., 2002; Skold *et al*., 2002; Svensson *et al*., 2003; Tammen *et al*., 2003; Wang *et al*., 2003; Anderle *et al*., 2004; Li *et al*., 2004; Radulovic *et al*., 2004; Wiener *et al*., 2004; Listgarten and Emili, 2005; Berg *et al*., 2006).

One of the major challenges in proteomics relates to looking for differences between samples belonging to different experimental groups (e.g. healthy/disease or control/treated). It is critical to minimize the variation between technical replicates, i.e. repeated analysis of the same sample (Venable and Yates, 2004), and to move the focus onto biological variation to allow for the sensitive detection of biologically relevant differences between the groups. Several factors are crucial, including sample quality, reproducibility of sample preparation, quality of the chromatography system used, and performance of the mass spectrometer. The reproducibility of the MS ion signal for technical replicates is investigated here and the reproducibility of protein identification is compared with it.

Because the outcome of an LC-MS experiment depends on many different variables, it is difficult to optimize the system by systematically optimizing individual variables. In this paper, examples are presented of how the 2D and 3D visualization approach of DeCyder™ MS Differential Analysis Software (DeCyder MS) (GE Healthcare), where retention time, precursor mass, and the topology of the intensity profile are co-visualized, can be used in combination with the matching of tandem mass spectra, to achieve a very high reproducibility within technical replicates.

DeCyder™ MS is a software intended for differential analysis of data from LC-MS experiments. It provides novel 2D and 3D visualizations of LC-MS data to allow for raw data quality assessment and interactive confirmation of results achieved using automated methods for peptide detection, charge state assignments, and peptide matching across multiple LC-MS experiments. Univariate statistical tools (Students *t* test and ANOVA) are available to identify significantly varying peptides among different groups of samples and variation patterns can be visualized in various graphs (A Kaplan, M Söderström, D Fenyö, H Pettersen, S Lindqvist, L Björkesten, unpublished data).

The technique described above was used to analyse protein abundance in samples which formed part of a study designed to identify genuine residents within plant organelles. In this study, a cellular extract from *Arabidopsis thaliana* non-photosynthetic callus cultures was prepared and a total membrane fraction applied to an iodixanol self-forming density gradient (Dunkley *et al*., 2004). Fractions from this gradient were analysed in a study not described here, in an attempt to match the distribution of protein of unknown location in the cell with that of known organelle markers. Here, four consecutive fractions from the lower end of this gradient, which was the site of enrichment of mitochondrial, plastid, and rough endoplasmic reticulum, were taken and proteins assessed in terms of reproducibility of technical replicates in LC-MS experiments where relatively complex fractions are analysed.

Materials and methods

Samples from organelle-enriched fractions of *Arabidopsis thaliana* were prepared (Dunkley *et al*., 2004). The fractions were digested using trypsin and analysed by one-dimensional LC-MS using an Ettan™ MDLC system (GE Healthcare) in high-throughput configuration directly connected to a Finnigan™ LTQ™ system (Thermo Electron). Samples were concentrated and desalted on RPC trap columns (Zorbax™ 300 SB C18, 0.3 mm×5 mm, Agilent Technologies), and the peptides were separated on a nano RPC column (Zorbax 300 SB C18, 0.075 mm×100 mm, Agilent Technologies) using a linear acetonitrile gradient from 0% to 48% ACN (GE Healthcare, 1% ACN increase min^{-1}). All buffers used for nano LC separation contained 0.1% formic acid (Fluka) as the ion pairing reagent. Full scan mass spectra were recorded in profile mode and tandem mass spectra in centroid mode. The peptides were identified using the information in the tandem mass spectra by searching against the *A. thaliana* proteome (Birney *et al*., 2004) using X!Tandem (Craig and Beavis, 2004) (Beavis Informatics) using an expectation value cut-off of 0.01. The expectation value is a measure of the statistical significance of the identification being true (Eriksson *et al*., 2000; Eriksson and Fenyo, 2002; Fenyo and Beavis, 2003). It is calculated by extrapolating the extreme value distribution of scores for randomly matching protein sequences observed at low scores.

The LC-MS data from the different samples was displayed as two-dimensional intensity maps with *m/z* and retention time on the two axes and a grey scale representing the intensity of a peak at a certain *m/z* and retention time using DeCyder™ MS Differential Analysis Software (DeCyder MS) (GE Healthcare) (A Kaplan, M Söderström, D Fenyö, H Pettersen, S Lindqvist, L Björkesten, unpublished data). DeCyder MS was also used to analyse the intensity maps in two steps. In the first step a dedicated image analysis algorithm was used to perform peptide detection, charge state assignment, and quantitation in the PepDetect module of the software. The detected peptides were indicated in the intensity maps by boxes and the MS/MS events were marked by crosses. The second step in the analysis was the matching of peptides falling within a user-defined mass and retention time interval in a comparison between different intensity maps

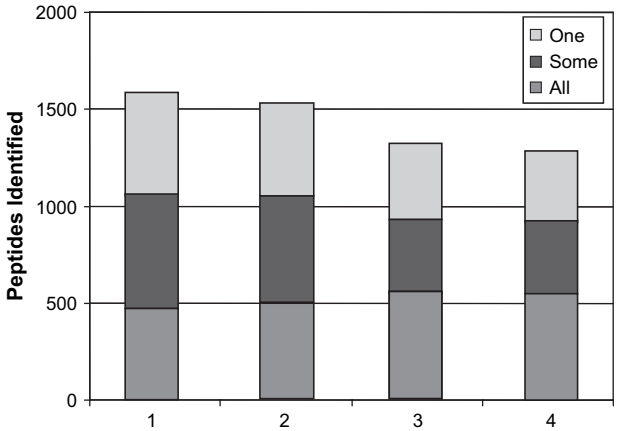

Fig. 1. The number of peptides identified using X!Tandem with expectation values less than 0.01 in all, some, and only one of the replicates from four different *Arabidopsis* samples. Samples 1 and 2 had five replicate analyses and samples 3 and 4 had four replicates.

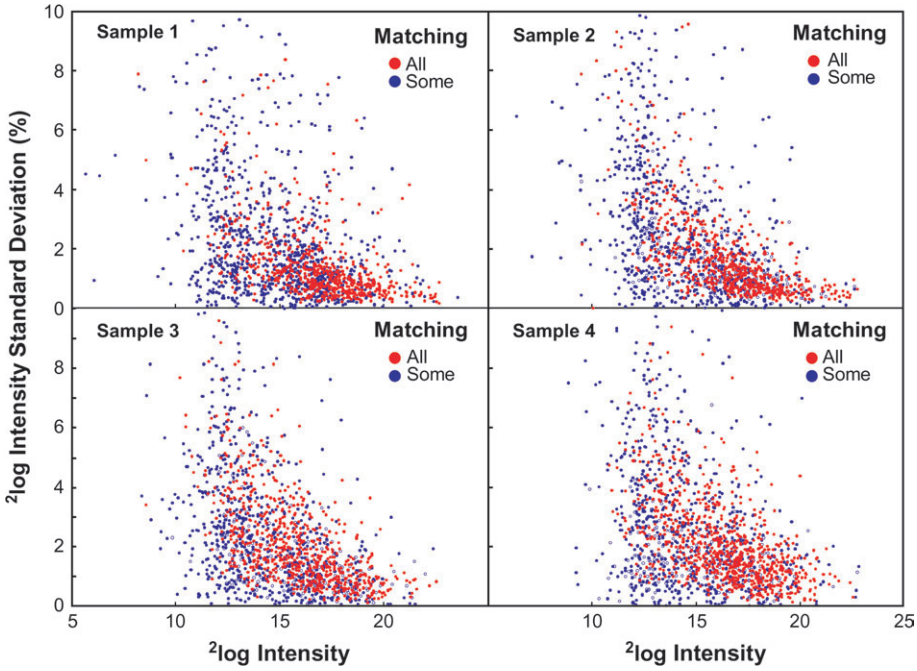

Fig. 2. The variation in intensity between different replicate runs is shown for four different samples. The peptides matching all replicate intensity maps are shown as red dots and the peptides matching some replicate intensity maps (e.g. 1–3) are shown as blue dots. The random variation in ^{2}log peak intensity between repeat analyses is in the range of a few percent, making it straightforward to compare repeat analyses by comparing intensity maps.

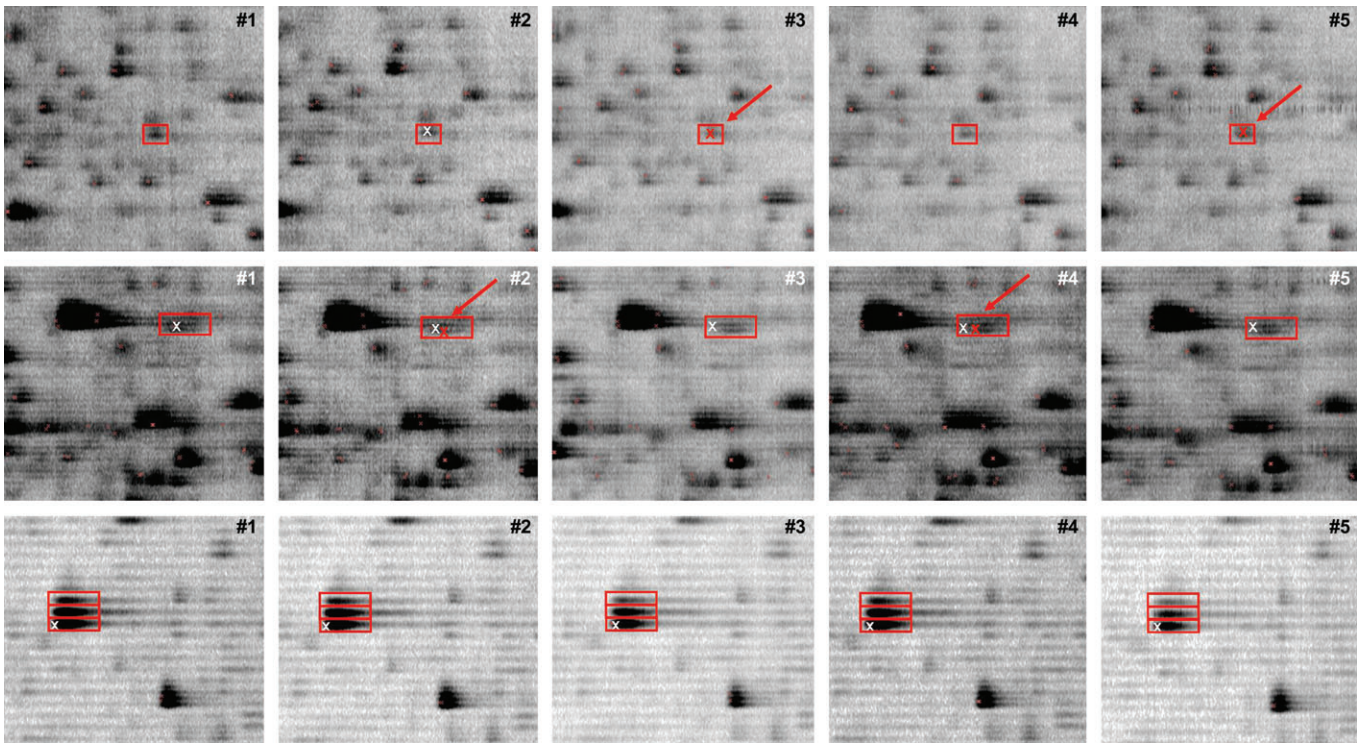

Fig. 3. Examples of intensity maps showing three peptides (From top: KGDLLLGDVAF, ASALIQHEWKPK, INAGLSFTK) from an *Arabidopsis* porin (Ensembl:At3g01280.1) for five replicate LC-MS runs clearly showing that these peptides are present in all runs. The tandem mass spectra that lead to a successful identification are indicated by red markers (expectation value, e <0.01) and the tandem mass spectra that did not lead to a successful identification are indicated by white markers. Note also the lack of tandem mass spectra (no marker) for some peptides.

from replicate analyses using the PepMatch module. MS/MS data corresponding to the detected peptides was exported and searched using X!Tandem and identification information imported back into DeCyder MS.

Results and discussion

Four different fractions isolated from a density gradient of a membrane preparation from *A. thaliana* were analysed by LC-MS in multiple technical replicates (four or five consecutive runs of the same sample). The data were evaluated in terms of the reproducibility between the different replicas.

The number of uniquely identified peptides found in one, some or all replicates is shown for each fraction in Fig. 1 and the peptide signal intensity reproducibility is illustrated in Fig. 2. The signal intensity distribution gives a first hint about the reproducibility of the LC-MS data and should be examined before further analysis. In this case, the intensity distributions indicate a good LC-MS data reproducibility in between replicas.

One of the identified proteins (porin, Ensemble: At3g01280.1) was taken as an example for closer evaluation. Figure 3 shows three peptides from At3g01280.1 that were automatically detected and matched across all replicates using DeCyder MS. Visual inspection of the LC-MS data, reveals that only two of the peptides have associated tandem mass spectra for all repeats. Furthermore, not all repeats with associated tandem mass spectra could be successfully identified (Table 1).

This demonstrates clearly that the software used for selecting ions for tandem mass spectrometric analysis and the identification algorithms are sensitive to small variations in peak intensity and tandem mass spectrum quality and therefore cause variations in the overall results of

Table 1. *The expectation value (*e*) for peptides with* e *<0.01 from an* Arabidopsis *porin (Ensembl:At3g01280.1) for five replicate LC-MS runs*

There is a variation in the expectation values for the same peptide in the different replicate analyses caused by variation in parent mass assignment and the varying quality of tandem mass spectra. For peptides that have an expectation value, e <0.01 in at least one of the replicates, the expectation value is shown in grey italics if $0.01<e<0.1$ for the other replicates. The results from the combined DeCyder™ MS and X!Tandem analysis is shown in the last column. Here, the peptides from the replicate runs were detected and matched using the DeCyder™ MS software before the tandem mass spectrometric information associated with the detected peptides was exported and searched with X!Tandem. This resulted in a more complete list of identified peptides as well as faster searches because of the smaller set of tandem mass spectra used in the search. The reason why peptides 6 and 23 did not show up in the DeCyder MS analysis is that they were not captured by the automated peptide detection algorithm in DeCyder MS. They were, however, clearly visible in the 2D visualizations and thereby available for manual inclusion.

	Peptides	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	DeCyder MS
1	AITSTGTKKGDLLLGDVAFQSR	2.00E-03	–	1.50E-04	3.80E-05	2.90E-05	*1.70E-02*
2	ASALIQHEWKPK	–	9.90E-03	–	8.40E-03	–	1.30E-03
3	DSTITVGTQHSLD	–	–	–	–	1.80E-05	7.10E-04
4	DSTITVGTQHSLDPLTSVK	1.80E-09	2.50E-09	1.80E-07	4.60E-09	6.10E-09	4.30E-10
5	EDLIASLTVNDK	1.20E-05	8.30E-06	1.10E-06	1.50E-06	7.30E-07	3.70E-08
6	EWKPKSFFTISGEVDTK	4.50E-03	*1.40E-02*	2.40E-03	*1.30E-02*	*8.10E-02*	–
7	FNTAVGAEVSHK	4.60E-06	1.20E-06	1.30E-04	7.70E-06	4.00E-06	7.10E-08
8	FSITTFSPAGVAITSTGTK	–	7.20E-12	5.90E-11	–	2.10E-04	3.30E-13
9	GDLLLGDVAFQSR	–	–	–	1.70E-05	–	3.90E-07
10	GDLLNASYYHIVNPLFN	1.30E-06	–	2.20E-06	3.70E-06	–	1.40E-07
11	GDLLNASYYHIVNPLFNTAVGAEVSHK	4.20E-11	3.810E-09	3.10E-07	7.50E-09	1.60E-07	1.50E-10
12	GPGLYTEIGK	4.20E-05	9.80E-08	3.60E-07	2.80E-07	2.10E-06	1.00E-07
13	GPGLYTEIGKK	8.20E-03	1.60E-03	1.60E-03	–	–	5.30E-04
14	GTQHSLDPLTSVK	–	–	1.30E-05	–	–	1.20E-04
15	HIVNPLFNTAVGAEVSHK	–	–	–	–	–	2.80E-12
16	IITHPNFNGNTLDNDIMLIK	4.10E-07	*1.60E-02*	*5.50E-02*	2.60E-03	*1.10E-02*	1.50E-10
17	INAGLSFTK	*2.00E-02*	*2.90E-02*	*6.50E-02*	*2.90E-02*	*3.10E-02*	4.40E-03
18	KGDLLLGDVAF	–	–	3.60E-04	–	3.50E-05	5.60E-05
19	KGDLLLGDVAFQSR	3.30E-12	4.00E-11	2.10E-11	9.50E-12	2.60E-11	9.90E-11
20	LGEHNIDVLEGNEQFIN	–	3.30E-08	5.20E-07	6.20E-08	2.40E-08	2.30E-08
21	LGEHNIDVLEGNEQFINAA	–	5.50E-09	6.50E-10	8.50E-09	4.90E-11	1.00E-10
22	LGEHNIDVLEGNEQFINAAK	1.10E-10	8.10E-07	6.40E-13	3.80E-11	2.70E-13	4.70E-14
23	LGEHNIDVLEGNEQFINAAKIITHPNFNGN	5.50E-07	1.40E-06	–	–	3.70E-06	–
24	LSSPATLNSR	*2.30E-02*	6.30E-03	*4.50E-02*	*4.40E-02*	*1.00E-02*	1.30E-03
25	PGLYTEIGK	4.00E-04	4.10E-03	3.80E-04	1.30E-03	2.20E-04	2.40E-04
26	SFFTISGEVDTK	1.30E-09	1.20E-09	1.30E-09	1.10E-10	2.20E-10	2.00E-09
27	SSPATLNSRVATVSLPR	*2.20E-02*	8.20E-03	*3.10E-02*	5.70E-03	–	*1.90E-02*
28	TVGTQHSLDPLTSVK	2.90E-07	–	1.00E-07	7.10E-08	3.00E-08	2.20E-07
29	VATVSLPR	2.50E-03	2.80E-03	5.40E-04	1.70E-03	1.40E-03	7.50E-04
30	VCTDSTFLITATVDEAAPGLR	3.40E-11	–	8.50E-13	2.00E-11	2.20E-11	4.00E-14
31	VELQYLHEY	2.80E-03	–	–	–	–	4.10E-03
32	VKGPGLYTEIGK	4.30E-05	4.20E-05	4.70E-04	8.50E-03	1.60E-04	3.30E-05
33	VNSAGIASALIQHEWKPK	6.00E-04	1.80E-08	3.80E-10	1.40E-09	9.20E-09	1.50E-09
		21	20	23	22	21	29

proteomics data. This sensitivity reduces the reproducibility of LC-MS data that could be obtained with modern instruments, and which could easily be seen using the visualization and matching tools provided in DeCyder MS. The data for the selected protein (Ensemble: At3g01280.1) are summarized in Table 1. Evaluation of data from all five technical replicates identify, in total, 33 different peptides of this protein. But in each single replicate not more than a maximum of 23 peptides have been identified, whereas using the strategy involving DeCyder MS for detection and matching of peptides between replicates resulted in 29 peptides being identified. This observation supports the need of comparing complete data sets on the basis of intensity maps to be able reproducibly to detect and assign the peptides observed

The visualization of LC-MS data as a two-dimensional intensity map resembles very much a 2D-PAGE image. This way of presenting of LC-MS data is much more intuitive to the human eye than the conventional way of inspecting the total ion chromatogram and individual mass spectra. Inspecting the intensity map can help to assess rapidly the overall quality of an LC-MS analysis.

The images from DeCyder MS can also be used to check the reproducibility and consistency of replicate sample analyses. Inconsistency in database search results from replicate analyses can be explained by inspecting the intensity maps showing the tandem mass spectrometric events. The differences between replicate analyses are due to the fact that tandem mass spectra are acquired at slightly different retention times and *m/z* values due to the variation in the intensity of peptides between replicate analyses. In some cases, tandem mass spectra corresponding to a peptide are acquired in some replicate analyses but not in others. Also, there will always be differences in the quality of the tandem mass spectra acquired causing variations in the scoring by the search engines. The peptide can, in most cases, still be detected and confirmed from its location in relation to neighbouring peptides in the intensity map.

Therefore, it is possible to achieve a very high reproducibility in proteomics experiments by visual inspection of the intensity maps, assuming the chromatographic separation is reproducible. It has been well established that images allow intuitive analysis and allow access to information that is otherwise not discernible by sequential examination of single spectra. The case of LC-MS is not an exception. Even though LC-MS image analysis is still in its infancy, the potential and advantages can now be shown by using DeCyder MS.

Conclusions

Small intensity changes between replicate analyses of the same sample cause variation in the data-dependent acquisition of tandem mass spectra and in the quality of the tandem mass spectra acquired, leading to variation in which peptides are identified by database searching. It is possible to assess and increase the reproducibility of repeat analysis by using the detection, matching, and 2D-visualization of DeCyder MS.

Acknowledgements

We thank Julie Howard, Tom Dunkley, and Kathryn Lilley, Cambridge Centre for Proteomics, University of Cambridge, UK, for supplying the *Arabidopsis* organelle-enriched fractions and for fruitful discussions.

References

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422,** 198–207.

Anderle M, Roy S, Lin H, Becker C, Joho K. 2004. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* **20,** 3575–3582.

Berg M, Parbel A, Pettersen H, Fenyo D, Björkesten L. 2006. Detection of artifacts and peptide modifications in LC-MS data using novel visualization software. *Rapid Communications in Mass Spectrometry* (in press).

Birney E, Andrews TD, Bevan P, *et al.* 2004. An overview of Ensembl. *Genome Research* **14,** 925–928.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467.

Dunkley TP, Dupree P, Watson RB, Lilley KS. 2004. The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in *Arabidopsis thaliana*. *Biochemical Society Transactions* **32,** 520–523.

Eriksson J, Chait BT, Fenyo D. 2000. A statistical basis for testing the significance of mass spectrometric protein identification results. *Analytical Chemistry* **72,** 999–1005.

Eriksson J, Fenyo D. 2002. A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis. *Proteomics* **2,** 262–270.

Fenyo D. 2000. Identifying the proteome: software tools. *Current Opinion in Biotechnology* **11,** 391–395.

Fenyo D, Beavis RC. 2003. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry* **75,** 768–774.

Heine G, Zucht HD, Schuhmann MU, Burger K, Jurgens M, Zumkeller M, Schneekloth CG, Hampel H, Schulz-Knappe P, Selle H. 2002. High-resolution peptide mapping of cerebrospinal fluid: a novel concept for diagnosis and research in central nervous system diseases. *Journal of Chromatography B, Analytical Technology Biomedical Life Sciences* **782,** 353–361.

Li XJ, Pedrioli PG, Eng J, Martin D, Yi EC, Lee H, Aebersold R. 2004. A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Analytical Chemistry* **76,** 3856–3860.

Listgarten J, Emili A. 2005. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular and Cell Proteomics* **4,** 419–434.

Palmblad M, Ramstrom M, Markides KE, Hakansson P, Bergquist J. 2002. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Analytical Chemistry* **74,** 5826–5830.

Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. 2004. Informatics platform for global proteomic profiling

and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Molecular and Cell Proteomics* **3,** 984–997.

Schulz-Knappe P, Zucht HD, Heine G, Jurgens M, Hess R, Schrader M. 2001. Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. *Combinatorial Chemistry and High Throughput Screening* **4,** 207–217.

Skold K, Svensson M, Kaplan A, Bjorkesten L, Astrom J, Andren PE. 2002. A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* **2,** 447–454.

Svensson M, Skold K, Svenningsson P, Andren PE. 2003. Peptidomics-based discovery of novel neuropeptides. *Journal of Proteome Research* **2,** 213–219.

Tammen H, Kreipe H, Hess R, Kellmann M, Lehmann U, Pich A, Lamping N, Schulz-Knappe P, Zucht HD, Lilischkis R. 2003. Expression profiling of breast cancer cells by differential peptide display. *Breast Cancer Research Treatment* **79,** 83–93.

Venable JD, Yates J III R. 2004. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Analytical Chemistry* **76,** 2928–2937.

Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. 2003. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry* **75,** 4818–4826.

Wiener MC, Sachs JR, Deyanova EG, Yates NA. 2004. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Analytical Chemistry* **76,** 6085–6096.