JOURNAL OF MASS SPECTROMETRY
J. Mass Spectrom. 2004; **39**: 770–780
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jms.650

JMS

Identification of low molecular weight proteins isolated by 2-D liquid separations

Kan Zhu,[1] Fred R. Miller,[2] Timothy J. Barder[3] and David M. Lubman[1]*

[1] Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055, USA
[2] Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, Michigan 48201, USA
[3] Eprogen, Inc., 8205 S. Cass Avenue, Suite 111, Darien, Illinois 60561, USA

Received 27 January 2004; Accepted 8 April 2004

Proteins with molecular mass (M_r) <20 kDa are often poorly separated in 2-D sodium dodecyl sulfate polyacrylamide gel electrophoresis. In addition, low-M_r proteins may not be readily identified using peptide mass fingerprinting (PMF) owing to the small number of peptides generated in tryptic digestion. In this work, we used a 2-D liquid separation method based on chromatofocusing and non-porous silica reversed-phase high-performance liquid chromatography to purify proteins for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric (MALDI-TOFMS) analysis and protein identification. Several proteins were identified using the PMF method where the result was supported using an accurate M_r value obtained from electrospray ionization TOFMS. However, many proteins were not identified owing to an insufficient number of peptides observed in the MALDI-TOF experiments. The small number of peptides detected in MALDI-TOFMS can result from internal fragmentation, the few arginines in its sequence and incomplete tryptic digestion. MALDI-QTOFMS/MS can be used to identify many of these proteins. The accurate experimental M_r and p*I* confirm identification and aid in identifying post-translational modifications such as truncations and acetylations. In some cases, high-quality MS/MS data obtained from the MALDI-QTOF spectrometer overcome preferential cleavages and result in protein identification. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: low molecular weight protein; peptide identification; matrix-assisted laser desorption/ionization; quadrupole time-of-flight; two-dimensional; liquid separation; breast cancer

INTRODUCTION

Peptide mass fingerprinting (PMF) has been widely applied in identifying proteins separated by two-dimensional (2-D) gel electrophoresis or other multi-dimensional chromatographic methods.[1] In this method, proteins are enzymatically digested and peptide masses are measured by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOFMS) and experimental masses are correlated with proteins in a database using various algorithms.[2] Some of these algorithms[1,3] rank proteins according to the decreasing number of matching peptides whereas other methods[4] use a probability-based algorithm. An elegant ranking method proposed by Pappin *et al*.[5] combines the frequency distribution of the matching peptides and provides a normalized probability related score, MOWSE. The success of all these methods depends on the number of experimental masses that match the theoretical masses. A larger number of matching peptides result in an improved search result.

The success of PMF relies on the accuracy of the mass measurement, the number of peptides matched, the sequence coverage achieved and the complexity of the protein. Currently, commercial MALDI-TOF mass spectrometers with delayed extraction and reflectron capabilities can easily achieve a mass accuracy of <25 ppm with internal calibration. However, the number of peptides produced by enzymatic digestion is size or molecular mass (M_r) dependent. In general, proteins with high M_r generate more peptides than those of low M_r. However, not all peptides that result from digestion can be detected in MALDI-TOFMS. In most MALDI-TOF instruments there is a useful m/z range of 700–4000, where below 700 the signals are masked by matrix peaks and above 4000 monoisotopic resolution is often not achieved. The number of peptides observed is also affected by the MALDI process. Usually the lysine-terminated peptides result in a lower signal compared with arginine-terminated peptides[6] owing to the lower proton affinity in the gas phase,[7] where many of these peptides remain undetected. Modifications, such as phosphorylation and glycosylation, also decrease the MALDI efficiency. Also, there is sample loss due to sample preparation, especially during in-gel digestion in 2-D gel separated proteins so that the number of peptides

*Correspondence to: David M. Lubman, Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055, USA.
E-mail: dmlubman@umich.edu
Contract/grant sponsor: National Institutes of Health;
Contract/grant number: R01 GM 49500.
Contract/grant sponsor: National Cancer Institute; Contract/grant numbers: R21CA83808; R01CA90503.
Contract/grant sponsor: Eprogen, Inc.

observed is considerably smaller than that expected. As a result, there may not be a sufficient number of peptides detected for unambiguous protein identification with the PMF method, especially for small proteins with $M_r < 20$ kDa.

Efforts to improve the results from PMF include increased mass accuracy,[8] sequence information obtained from isotopically labeled peptides,[9–12] specificity gained using different enzymatic digestions[13] and increased efficiency in MALDI by converting lysine-terminated peptides to arginine-terminated peptides through a guanidination reaction.[14,15] These steps require chemical modification or high-accuracy mass spectrometers. An alternative to PMF in protein identification is tandem mass spectrometry (MS/MS). The advantage of the method is that partial sequence of a single peptide is often adequate for high-confidence protein identification.

In this study, low-mass proteins were purified from breast cancer cell whole cell lysates by 2-D liquid-phase separation. Protein identification for low-M_r proteins was compared using PMF, modified PMF and the MALDI-QTOF method. The MALDI-QTOF technique proved to be the most effective method for identifying low-mass proteins when coupled with the precise M_r obtained for intact proteins from the liquid-phase separations. The accurate M_r confirms the protein identification achieved by using the MS/MS method. In many cases, the M_r information is necessary for correct protein identification, especially in human cells in which there may be several modified isoforms of a protein present and where there is often sequence homology between proteins. Nevertheless, the use of all these techniques aids in identification in these often modified human proteins.

EXPERIMENTAL

MCF10CA1d.cl1 cell preparation and lysis

Fully malignant human breast cancer cells, MCF10CAld clone 1 (CA1d) cells,[16] were grown in monolayer on plastic in DMEM/F12 medium supplemented with 5% horse serum, 10 mM HEPES (*N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid). Harvesting was performed in the log phase (~75–80% confluence). The cells were gently washed with sterile PBS buffer before they were scraped with a rubber policeman and stored at −80 °C, after the growth medium was aspirated.

A 1.0 ml volume of lysis buffer containing 6 M urea, 2 M thiourea, 0.5% (w/v) *n*-octyl β-D-glucopyranoside (OG), 50 mM dithiothreitol (DTT), 2 mM tris(2-carboxyethyl)phosphine (TCEP), 1 µl of protease inhibitor (Sigma, St. Louis, MO, USA) and 10% (v/v) glycerol was added to cells prior to vortex mixing for 2 min. The mixture was then left at room temperature for 1 h. In order to eliminate cell debris, sample and buffer mixture were centrifuged at 20 000 *g* for 20 min. The supernatant was collected and the protein content was determined using a Bradford-based assay (Bio-Rad, Hercules, CA, USA).

Chromatofocusing (CF)

As described in previous work,[17] lysis buffer in the whole cell lysate was replaced by the start buffer, which contained 25 mM bis–tris propane (Sigma) and 6 M urea. The pH was adjusted to 7.4 with saturated imminodiacetic acid (Sigma) using a PD-10 (Amersham Pharmacia, Piscataway, NJ, USA) according to the manual prior to the sample loading. About 7 mg of protein extracted from CAld cells was loaded on to the CF column followed by elution at 0.2 ml min^{-1} using as elution buffer a mixture of 10% (v/v) PolyBuffer 74 (Amersham Pharmacia) and 6 M urea with a pH of 4.0. Saturated imminodiacetic acid was applied for pH adjustment if necessary. On-line pH measurement was performed off-column before fraction collection using a pH electrode (Lazar Research, Los Angeles, CA, USA) and the separation was monitored at 280 nm using a Beckman Model 126 UV detector (Beckman-Coulter, Fullerton, CA, USA). Eluate from CF separation was collected from pH 7.4–4.2 at 0.2 pH unit intervals. The collection process was controlled using software written in-house. All fractions collected were stored in a dry-ice box after collection. Samples were stored in a −80 °C freezer.

Non-porous (NPS) reversed-phase high-performance liquid chromatography (RP-HPLC)

The NPS RP-HPLC separation was performed at a flow-rate of 0.5 ml min^{-1} on an ODS IIIE column (33 × 4.6 mm i.d.) packed with 1.5 µm non-porous silica beads derivatized with C_{18} (Eprogen, Darien, IL, USA). Each CF fraction contained more than 50 µg of protein, all of which was loaded on the NPS RP column. During separation, the column was maintained at 65 °C with a Timberline (Boulder, CO, USA) column heater in order to improve the resolution and speed of the separation. Trifluoroacetic acid (TFA) (0.1% (v/v)) was added to both mobile phase A (deionized water) and mobile phase B (acetonitrile). The gradient profile used was as follows: (1) 5 to 15% B in 1 min; (2) 15 to 25% B in 2 min; (3) 25 to 31% B in 3 min; (4) 31 to 41% B in 10 min; (5) 41 to 47% B in 3 min; (6) 47 to 67% B in 4 min; (7) 67 to 100% B in 1 min; (8) 100% B for 2 min; (9) 100 to 5% B in 1 min. The acetonitrile was 99.93% HPLC grade (Sigma) and the TFA was taken from 1 ml sealed ampules (Sigma). The HPLC system was a Beckman Model 127 and the separation was monitored at 214 nm using a Model 166 detector. The proteins separated by NPS HPLC were collected in 1.5 ml tubes using a Beckman SC-100 fraction collector controlled by in-house software. Purified samples were stored in dry-ice after collection.

NPS RP-HPLC/electrospray ionization (ESI) TOFMS

In order to obtain the M_r for intact proteins, the CF fraction with the same pH value from another CF separation was reseparated by NPS RP-HPLC where the eluate from NPS RP-HPLC was monitored on-line using ESI-TOFMS (LCT, Micromass, Manchester, UK). The separation was performed under the same conditions as in the previous section. However, in addition to 0.1% TFA, 0.3% formic acid (Sigma) was added to both mobile phases to improve the ESI efficiency. The eluent was introduced into the ESI-TOF system at a flow-rate of 200 µl min^{-1}. The capillary voltage for electrospray was set at 3200 V, sample cone at

40 V, extraction cone at 3 V and reflection lens at 750 V. Desolvation was accelerated by maintaining the desolvation temperature at 300 °C and the source temperature at 120 °C. The nitrogen gas flow was controlled at 650 l h^{-1}. During the separation, one mass spectrum was acquired per second. The intact molecular mass value was obtained by deconvoluting the combined ESI spectra of the protein with MaxEnt 1 software (Micromass).

Peptide preparation for MS analysis

Prior to tryptic digestion, purified proteins were dried down to 20 µl using a CentriVap concentrator (Labconco, Kansas City, MO, USA) followed by the addition of 20 µl of 100 mM ammonium bicarbonate (Sigma) to adjust the solution pH to ~7.8. A 0.5 µg amount of L-1-tosylamido-2-phenylethyl chloromethyl ketone (TPCK) modified sequencing grade trypsin (Promega, Madison, WI, USA) was added prior to vortex mixing. The mixture was stored in a warm room with a controlled temperature of 37 °C for 24 h where the samples were shielded from light. After digestion had been halted by adding 1 µl of 10% (v/v) TFA, the tryptic digest mixture was concentrated to 5 µl using a ZipTip (Millipore, Billerica, MA, USA). In order to prevent methionine oxidation, light exposure was minimized by storing the peptide mixture in a dark-box before and after concentration.

MALDI-TOFMS

Sample spotting was performed by layering 1 µl of matrix on top of 1 µl of sample. The spot was allowed to dry in air but without light exposure. The MALDI matrix was prepared by diluting saturated α-cyano hydroxy cinnamic acid (α-CHCA) (Sigma) solution made in 50% (v/v) acetonitrile and 1% (v/v) TFA with the same solution in a 1:4 (v/v) ratio. Internal standards were prepared as 1 mg ml^{-1} angiotensin I, ACTH 1–17 and ACTH 18–39 (Sigma) and were further diluted 100-fold with deionized water. Volumes of 13, 21 and 25 µl were taken from each of the diluted standard solutions and mixed with matrix to produce 1 ml of solution containing 50 fmol of each standard per spot.

Peptide masses were measured on a Micromass Tof-Spec2E system (Micromass/Waters, Milford, MA, USA) with delayed extraction in the reflectron mode using a nitrogen laser (337 nm). Peptide mass spectra were internally calibrated using the three peaks from the internal standards resulting in a mass accuracy of 50 ppm or less. The calibrated spectra were processed using PeptideAuto (Micromass MassLynx application) to obtain experimental masses that were submitted to MS-Fit (http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm) to search the SwissProt database for protein identity. The following parameters were used for database searching: maximum number of two missed cleavages, unmodified cysteine, *Homo sapiens* for the species, no restriction on the pH range, monoisotopic mass. In terms of modification, acetylation was allowed. However, only when the methionine was oxidized was 'methionine oxidation' included in the database search.

MALDI plate light exposure

Following preparation and analysis under shielded conditions, MALDI target plates were exposed to direct sunlight for ~20 h for 3 days by placing them on a sunny windowsill. After the light exposure, a second MALDI-TOF spectrum was acquired. For comparison, ACTH 1–17 standard was spotted on the MALDI plate and MALDI-TOF spectra were acquired before and after the storage for 3 days in a box shielded from light. Oxidation was not induced under these conditions. No additional steps were required to study methionine oxidation.

MALDI-QTOF

After analyzing the MALDI-TOF data, the same MALDI plate was reanalyzed using a Micromass MALDI Q-TOF Ultima system (Micromass/Waters) for selected peptides. According to the MALDI-TOF data, three peptides were selected from each spot for fragmentation. For parent ions smaller than 900 and larger than 2700 Da, the collision energy was set for 40 and 160 eV, respectively. Higher collision energy was applied for parent ions of higher M_r where the correlation is linear. The tandem mass spectra were acquired from m/z 50 to 50 + precursor ion. The nitrogen laser (337 nm) was scanned over each sample spot in a raster pattern for a total of 1680 s, 560 s for each parent ion. The parent ion window was set at 5 Da. Spectra of ACTH 1–17 were acquired for 60 s after completing MS/MS on each sample spot. Product ion calibration was achieved by applying the lock mass calibration obtained from ACTH 1–17. The calibrated tandem mass spectrum was subtracted, smoothed and centered before QTOF transformation, a process that transforms a spectrum with multiply charged ions into a spectrum of singly charged monoisotopic ions. Fragment ions were selected from the tandem mass spectrum manually and searched against SwissProt using MS-Fit and Mascot for the peptide sequence and methionine oxidation.

RESULTS AND DISCUSSION

Low molecular mass proteins identified by the PMF method

Low-M_r proteins are not well separated by current tris-glycine gel electrophoresis method compared with medium- and large-sized proteins. Partial improvements can be achieved by using tris-trycine gels.[18] However, post-separation sample loss during procedures for MS analysis by in-gel digestion or by blotting methods is significant. Keller *et al*. reported improved analysis of low-M_r proteins by accumulating proteins from multiple LC separations using a nanoliter sample handling technique.[19] Alternatively, the 2-D liquid separation method using CF followed by NPS RP-HPLC provides high-resolution separation of low-M_r proteins. In addition, the sample recovery for small proteins is often close to 90%. Losses are also minimal in the digestion process following collection in the liquid phase.

An accurate M_r value for intact proteins provides essential information that aids protein identification and characterization, such as post-translational modifications.[20,21] In order to obtain the M_r for intact proteins, on-line HPLC was performed for each of the fractions collected from CF. The HPLC conditions, such as gradient, column and temperature, in each LC/MS experiment were the same as those in

the offline collection, which was monitored by a UV detector at 214 nm. Hence the retention time of the LC separation can be used to correlate the M_r obtained from the on-line HPLC/MS experiments and the protein identity obtained from the analysis of the tryptic peptides of the collected protein that elute at the same retention time. Two of the ESI-TOF spectra acquired using on-line NPS HPLC/MS are shown in Fig. 1(A) and 1(B). The M_r values of intact proteins were produced by deconvoluting Fig. 1(A) and 1(B) using MaxEnt 1 software and are shown in Fig. 1(C) and 1(D). The intensity of the well-defined multiply-charged peaks suggest the abundance of these proteins.

Protein digestion was performed using trypsin and a peptide mass map obtained using MALDI-TOFMS. When the peptide masses for each of the purified proteins were submitted to SwissProt using MS-Fit, the first four proteins in Table 1 were identified. In all cases, $\geq$5 peptides and more than 34% sequence coverage were achieved with galectin having the highest sequence coverage of 70%. The MALDI-TOF spectrum of galectin is shown in Fig. 2. The identification was based on the suggestion that at least five matched peptides with mass accuracy within 50 ppm and sequence coverage of at least 15% are required to obtain unambiguous identification using the PMF method.[22] All identities were confirmed by the M_r values obtained from the ESI-TOF experiments. The discrepancy between the experimental and the theoretical M_r values may result from as yet unidentified post-translational modifications.

Despite the success of PMF in protein identification, the method still has many limitations. The method relies on the number of detected peptides where a number of factors can prevent peptides from being observed. These may include internal fragmentation, a weak response for some of the lysine-terminated peptides, non-specific and incomplete enzymatic digestion and modifications, which all result in insufficient peptides observed for unambiguous protein identification using PMF. In Table 1, all identified proteins have close to 10 theoretical tryptic peptides in the 'working range' of MALDI-TOFMS. In contrast, the proteins that were not identified have a smaller number of theoretical peptides in the range, where all have $\leq$3 peptides recorded. In the case of ubiquitin, there are only four theoretical peptides in the range, all of which were detected. Without the intact M_r value and peptide sequencing information, it still does not meet the criteria for unambiguous identification with PMF.

Protein identification enhancement with methionine oxidation as a sequence tag

Improvements in the PMF method have been explored extensively owing to the continuous increase in protein database size.[9] An increase in mass accuracy significantly enhances PMF search results and eliminates false positives.[10,23] A single mass detected by Fourier transform ion cyclotron resonance with mass accuracy of 0.1 ppm can achieve unambiguous identification for yeast proteins with the constraints of cysteine, enzyme and M_r of a protein.[24,25] Other methods for enhancing protein identification involve peptide sequence information from isotope and mass tags. Among them, alkylation of cysteine,[26] *in vitro* isotopically labeling of methionine, serine and tyrosine[11,27] and *in vivo* labeling,[28] such as leucine,[12] have been studied. Further modifications can be achieved by guanidination of lysine on the C-terminus of the tryptic peptides and Edman-type phenylthiocarbamoylation (PTC) of the N-terminus.[14,29,30] By converting lysine to homoarginine, the response of lysine-terminated peptides is enhanced. As a result, the lysine residues in the sequence are revealed and the lysine-terminated peptides are observed.

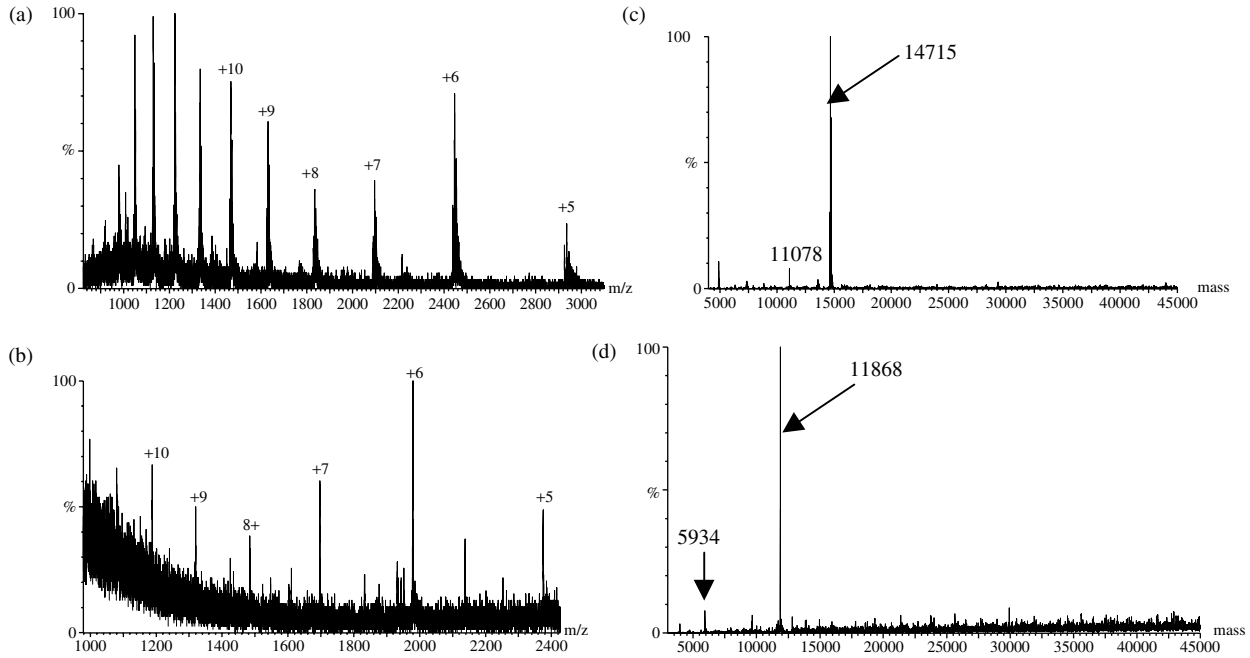

Figure 1. ESI-TOF spectrum of galectin (A) and thioredoxin (B). Labels on peaks represent the charge state. Their corresponding deconvoluted spectra are shown in (C) and (D). Numbers represent the molecular mass of intact proteins. 5934 in (D) is a harmonic peak, an artifact generated during deconvolution using MaxEnt 1 software.

Table 1. The number of peptides theoretically digested with trypsin vs the number of the observed peptides on MS

Protein	Theoretical M_r/pI	Experimental M_r	Theoretical tryptic digest (with 0 miscleavage)	Theoretical tryptic digest >700 Da	Detected on MS	Sequence coverage (%)
Galectin (P09382)	14 716/5.3	14 715	13	9	7	70
Stathmin (P16949)	17 303/5.8	17 311	33	10	6	36
Glia maturation factor beta (P17774)	16 713/5.2	16 784	22	9	5	47
ATP synthase D chain, mitochondrial (O75947)	18 360/5.2	18 403	21	10	5	34
Ubiquitin (P02248)	8564/6.6	8564	12	4	4	47
TCTP (P13693)	19 596/4.8	19 607	19	8	3	20
Centrin 2 (P41208)	19 738/4.9	19 807	31	9	3	29
Barrier to autointegration factor (O75531)	10 059/5.8	10 048	12	6	2	40
Cytochrome *c* oxidase polypeptide VA (P20674)	12 436/5.0	12 443	10	7	2	20
Thioredoxin (Q99757)	18 383/8.5[a]	11 868	18	9	2	37
40S ribosomal protein S21 (P35265)	9111/8.7	9154	13	5	1	18
ATP synthase coupling factor 6, mitochondrial precursor (P18859)	12 588/9.5[b]	8962	13	4	1	25
Cystatin B, human (P04080)	11 140/7.0	11 182	11	6	2	33
Thymosin β_{10} (P13472)	4894/5.3	4936	9	3	1	32
Ribosomal protein p2 (P05387)	11 665/4.4	11 662	10	6	3	28
Replicatrion protein A3 (P35244)	13 569/5.0	13 425	11	7	2	27
U6 snRNA-associated Sm-like protein LSm7 (Q9UK45)	11 602/5.1	11 606	13	7	1	17

[a] The theoretical M_r/pI of thioredoxin (mitochondrial) are 18383/8.5, but the truncated form has theoretical values of 11 867/4.9.
[b] The theoretical M_r/pI of ATP synthase coupling factor 6 (mitochondrial) are 12 588/9.5, but the truncated form has theoretical values of 8960/5.5.

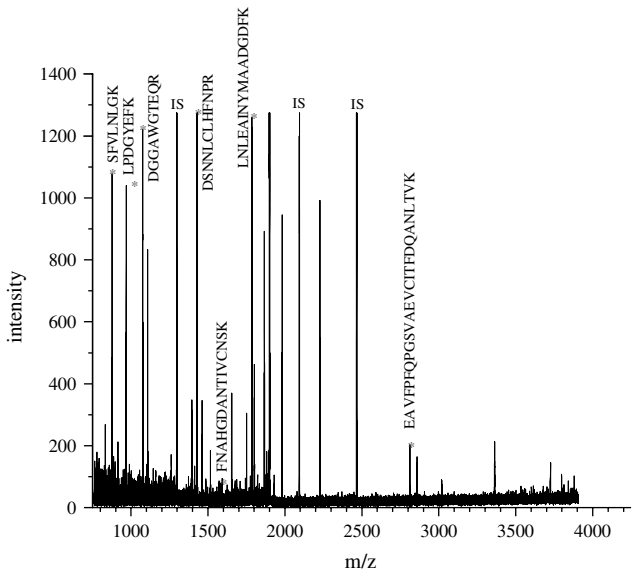

Figure 2. Galectin 1 MALDI-TOF spectrum.

Improved identification of proteins in PMF can also be achieved using methionine oxidation as a tag. This can be induced by a number of different methods, most notably using hydrogen peroxide or light. Methionine residues in methionine-containing proteins are not oxidized to sulfoxide during the 2-D liquid separation. If the methionine-containing peptides on the MALDI target plate are exposed to light for 3 days, new peaks due to the methionine oxidation are observed. The oxidation of methionine to sulfoxide results in peaks 16 Da higher than the original peak, a mass tag indicating the presence of methionine residues. As shown in Fig. 3, doublet peaks 16 Da apart indicate one Met in the peptide and triplet peaks 16 Da apart indicate two Mets in the peptide sequence. When the methionine sequence information was included in the database search, the result was improved, as suggested by the increase in MOWSE score as shown in Table 2. TCTP and centrin 2 are identified with confidence with the support from the experimental M_r for the intact protein, but not for barrier-to-autointegration factor and cytochrome *c* oxidase polypeptide VA because only two peptides were matched. Another problem for the method is that it is not broadly applicable since only methionine-containing peptides contribute to the enhancement.

Identification with MALDI-QTOF

General identification with MALDI-QTOF

Another protein identification method is based on peptide sequencing. Collision-induced dissociation (CID) amide bond fragmentation along the length of peptide generates b- and y-ions, which reveal the peptide sequence. Protein identification is achieved by searching the protein and nucleotide databases with product ions recorded in this experiment.[23,31,32] The partial sequence information obtained from MS/MS has more discriminating power than the PMF and may allow the identification of proteins based on a single peptide. The advantage is that small proteins may be identified if the peptide signal and its fragment signals are of sufficient intensity. The MALDI-QTOF system used in this work provides high resolution and the advantage of sharing the same sample preparation method with MALDI-TOFMS. A spot analyzed by MALDI-TOF can be reanalyzed

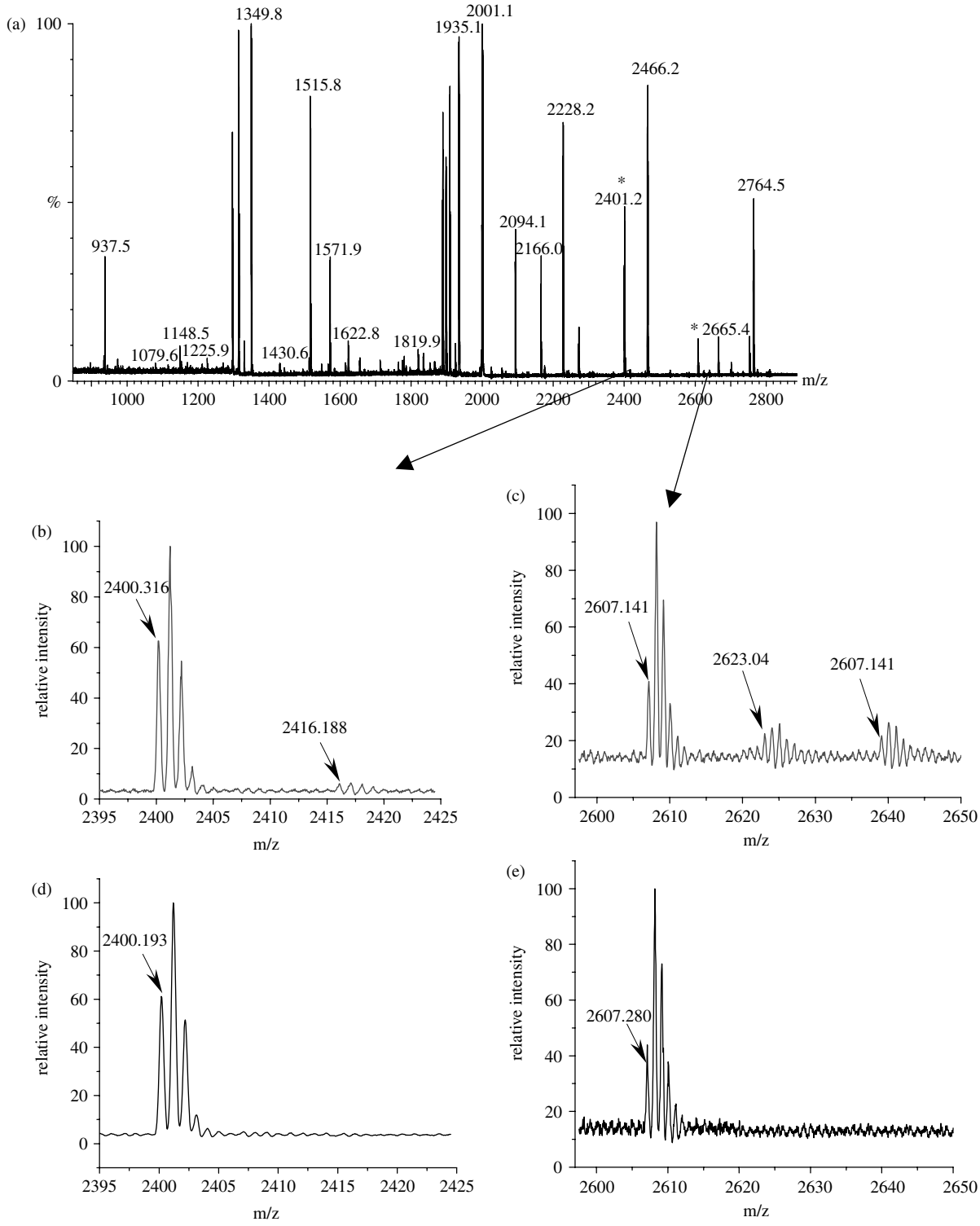

Figure 3. (A) MALDI-TOF spectrum of TCTP (P13693), GRP78, pyruvate kinase (P14786) and barrier-to-autointegration factor (O75531) mixtures. Peaks resulting from methionine oxidation were observed in the zoom-in spectra for barrier-to-autointegration factor (B) and TCTP (C) after light exposure for 3 days; their corresponding spectra prior to light exposure are shown in (D) and (E), respectively.

by MALDI-QTOF after selecting the peaks that need to be further analyzed. The quality of MS/MS data obtained and the effectiveness of using MALDI-QTOF in protein identification have been reported.[33–35]

Two of the MALDI-QTOF tandem mass spectra for the singly charged peptides 1429 (DSNNLCLHFNPR) and 1784 (LNLEAINYMAADGDFK) from galectin 1 are shown in Fig. 4. Both spectra show abundant b- and y-ions and are rich in sequence information. No dramatic difference in the abundance of product ions was observed, which indicates equal probability for amide bond cleavage across the backbone in CID. In the LNLEAINYMAADGDFK case, as

Table 2. Database search enhancement with Met oxidation

Protein	Observed Met-containing peptides	Database search result					
		Without Met oxidation			With Met oxidation		
		MOWSE score	Peptides matched	Coverage	MOWSE score	Peptides matched	Coverage
TCTP	DLISHDE<u>M</u>FSDIYKIR NYQFFIGEN<u>M</u>NPDG<u>M</u>VALLDYR	876	3	20	7.1×10^5	6	20
Centrin 2	<u>M</u>NFGDFLTV<u>M</u>TQK	1392	3	29	5.6×10^4	5	29
Barrier-to-autointegration factor	DFVAEP<u>M</u>GEKPVGSLAGIGEVLGK	254	2	40	2802	3	40
Cytochrome *c* oxidase polypeptide VA	GINTLVTYD<u>M</u>VPEPK	252	2	20	2040	3	20

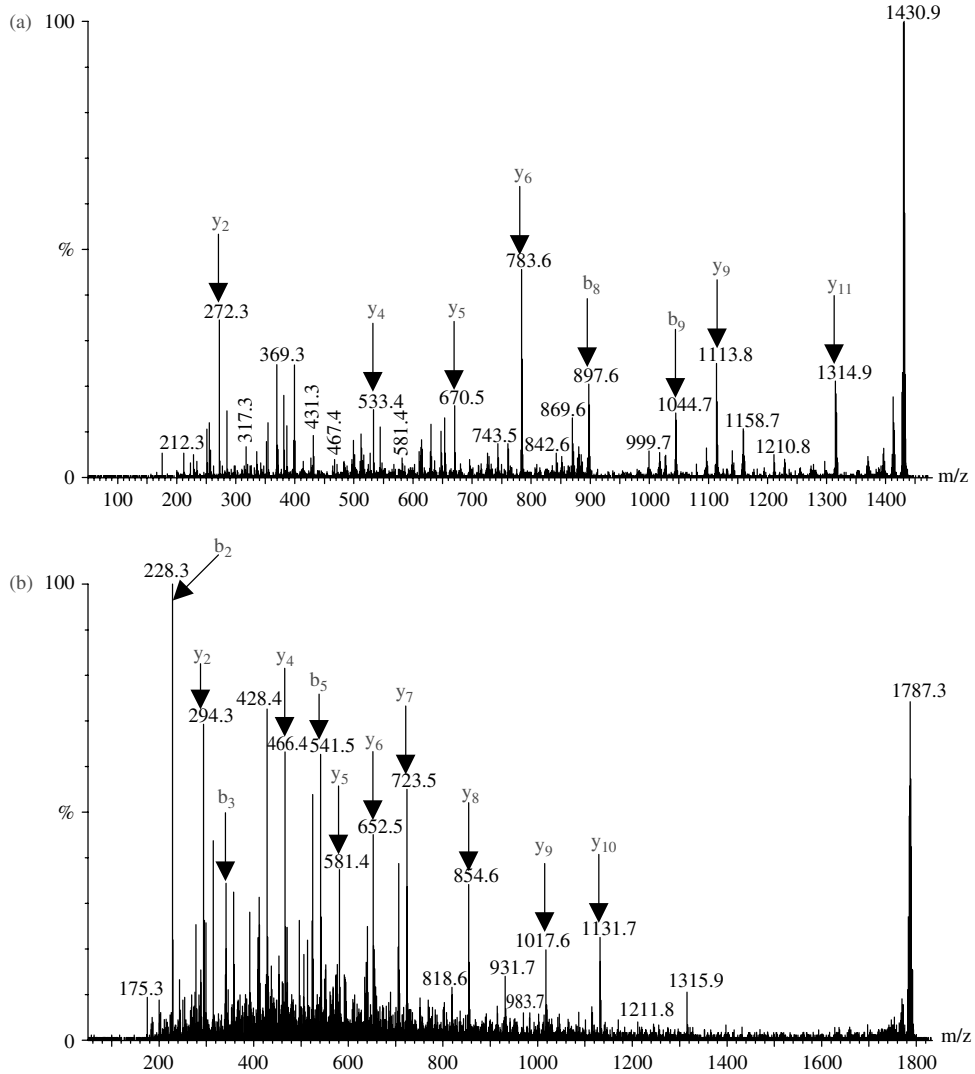

Figure 4. MALDI-QTOF tandem mass spectra of tryptic digest peptides from galectin 1. (A) DSNNLCLHFNPR (1429); (B) LNLEAINYMAADGDFK (1784).

in Fig. 4(B), multiple peaks from y_4 to y_{10} clearly sequenced NYMAAD. When product ions for each peptide and parent ion were used to search against SwissProt using MASCOT and MS-Tag, galectin 1 was unambiguously identified. The accurate M_r suggests that no modifications occur.

Thioredoxin and 40S ribosomal protein S21 are relatively low-abundance proteins in the breast cancer sample. Only two peptides were detected in MALDI-TOFMS for thioredoxin and one for 40S ribosomal protein S21. Nevertheless, they were identified by MS/MS data acquired by

the MALDI-QTOF spectrometer based on 1751 (FVGIKDEDQLEAFLK) and 1796 (MQNDAGEFVDLYVPR + acetyl N-terminus), respectively. The theoretical M_r and pI values from the database for thioredoxin are 18 383 and 8.5, respectively, which are significantly higher than the experimental M_r of 11 868 Da and pI 4.8–5.0. The first 59 amino acids were truncated when entering mitochondria, so that the accurate experimental M_r and pI values suggest that it is present in its mature form. When the truncation is taken into account and the theoretical M_r and pI are recalculated without the first 59 amino acids in its sequence, the experimental and the theoretical values match perfectly. In the case of the 40S ribosomal protein S21, the intact M_r obtained also supports the protein identification and the acetylation at the N-terminus. Truncated ATP synthase coupling factor 6 has an M_r of 8962, which only results in four peptides above 700. As in the case of ubiquitin, it cannot be identified by PMF alone, even if all four peptides are detected. However, fragmentation of 2081 (QTSGGPVDASSEYQQELER) by QTOF reveals its identity, which was also confirmed by an accurate experimental M_r.

Internal fragmentation

Internal fragmentation may result in an insufficient number of peptides for protein identification using PMF owing to the mismatch of the observed fragment masses with the theoretical masses in the database. Such internal fragmentation may result from sample preparation or unspecific digestion. A QTOF tandem mass spectrum of IGENMNPDGMVALLDYR, which is an internal fragment of NYQFFIGENMNPDGMVALLDYR from TCTP, is shown in Fig. 5. Various b and y ions were detected. On submitting the MS/MS data for protein identification without enzyme specification, TCTP was identified, whereas it remains unmatched in the PMF method.

Proteins with few arginines in their sequence

In another case, some proteins may contain few arginines in their sequence. The result is that the number of peptide masses acquired by MALDI-TOFMS is not sufficient for an unambiguous identification. Cystatin B only has two arginines present in its sequence, as shown in Fig. 6(A), so that only one mass was recorded in MALDI-TOFMS. However, the high-quality QTOF tandem mass spectrum shown in Fig. 7 clearly reveals the sequence VHVGDEDFVHLR, which brings cystatin B to the top of the list in a MASCOT search. The experimental M_r of the intact protein is 42 Da higher than its theoretical M_r, which indicates acetylation of the protein. Acetylation is a probable modification at its N-terminal site, although no further result was obtained to support the result. On the other hand, thymosin β_{10} contains only one arginine in its sequence, as shown in Fig. 6(B). Its identification was achieved by fragmentation of 1565 (ADKPDMGEIASFDK + acetyl) using the QTOF spectrometer. In an extreme case, thymosin β_4 contains no arginine in its sequence so that only two weak peaks corresponding to two peptides were recorded. These were not of sufficient intensity for the MALDI-QTOF experiment, although the abundance of this protein is high enough for capillary electrophoresis (CE)/MS and CE/MS/MS analysis for its identification.[20]

Incomplete tryptic digestion

Other than unspecific tryptic digestion and internal fragmentation, there may be multiple miscleavages in tryptic digestion. When only one miscleavage was allowed in the database search for the acquired masses from a tryptic digest of cytochrome *c* oxidase polypeptide VA, only two peptides, 2664 and 1356 (VIQELRPTLNELGISTPEELGLDK and NKPDIDAWELR), were matched, where one of the two peptides has a miscleavage. The sequence of both peptides

1 MMCGAPSATQ PATAETQHIA DQVRSQLEEK ENKKFPVFKA VSFKSQVVAG TNYFIKVHVG 60 A

61 DEDFVHLRVF QSLPHENKPL TLSNYQTNKA KHDELTYF

1 ADKPDMGEIA SFDKAKLKKT ETQEKNTLPT KETIEQEKRS EIS B

Figure 6. (A) Cystatin B sequence. Two arginines present in its sequence. (B) Thymosin β_{10} sequence. Only one arginine present in its sequence. R residues are underlined.

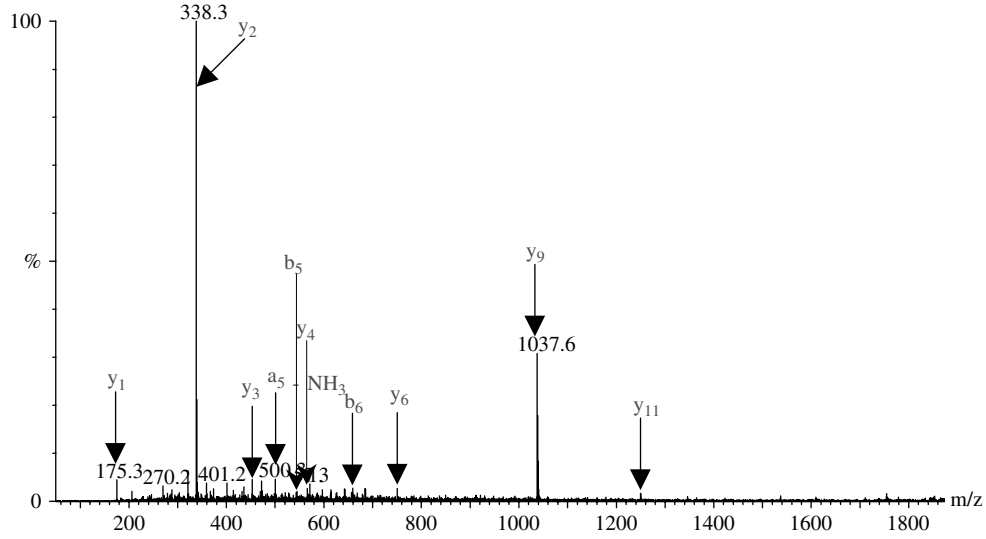

Figure 5. MALDI-QTOF tandem mass spectrum for 1908 (IGENMNPDGMVALLDYR) from TCTP, which is an internal fragment of NYQFF**IGENMNPDGMVALLDYR** (2607).

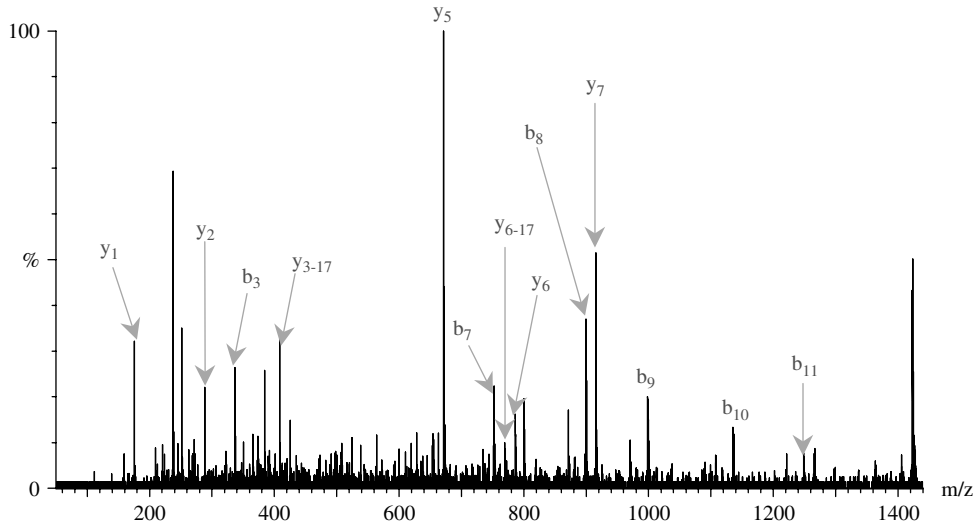

Figure 7. MALDI-QTOF tandem mass spectrum of VHVGDEDFVHLR from cystatin B.

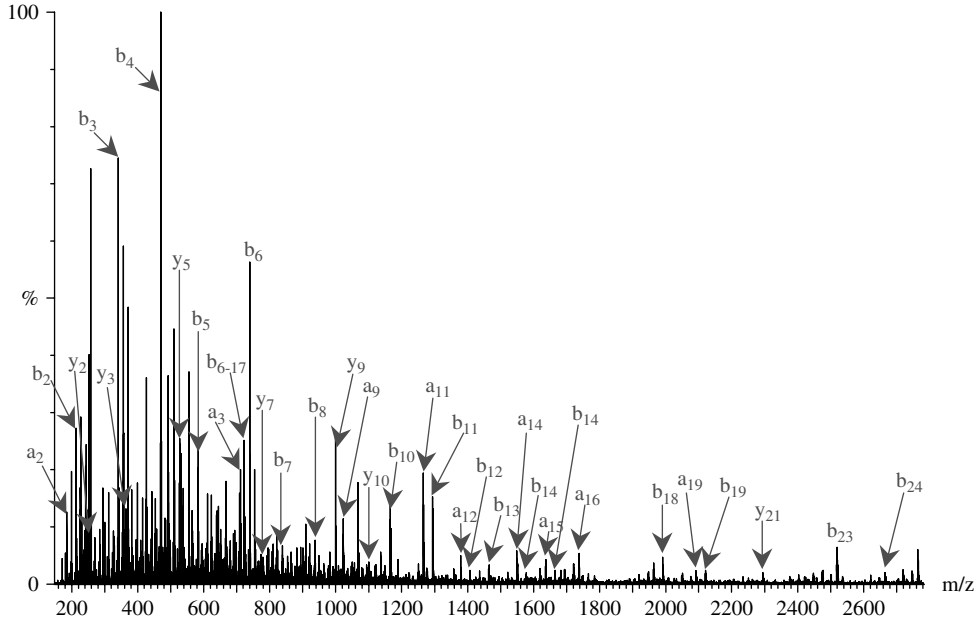

Figure 8. MALDI-QTOF tandem mass spectrum of VIQELRPTLNELGISTPEELGLDKV from cytochrome *c* oxidase polypeptide VA.

was confirmed by MS/MS experiments. However, when two miscleavages were allowed, another peptide, 2763 (VIQELRPTLNELGISTPEELGLDKV), which has one R and one K in its sequence, was recovered. The tandem mass spectrum of the peptide is shown in Fig. 8. The appearance of 2664 and 2763 at the same time suggests incomplete tryptic digestion where the two peptides are different and 2763 has a valine at its C-terminus. The incomplete digestion is due to the slow digestion rate of trypsin when P or D is adjacent to R and K. However, even when incomplete digestion results in unassigned peptides, QTOFMS can still recover correct protein identification.

Preferential cleavage

One problem with the MALDI-based MS/MS technique is that the dissociation of protonated peptide is not always evenly distributed across the peptide backbone. The singly charged peptide generated in MALDI favors cleavages C-terminal to aspartic acid (D) and glutamic acid (E) if the C-terminus is arginine.[36,37] This is clearly shown in Figs 5 and 9, where y_2 and y_9 of IGENMNPDGMVALLDYR and y_3, y_7 and y_{15} of GFDPLLNLVLDGTIEYMR dominate, respectively. The disadvantage is that a contiguous series of backbone cleavage sequence ions may not be obtained in a MALDI tandem mass spectrum. Given that current popular search algorithms are based on the assumption that amide bonds are cleaved with the same probability across the backbone of the protonated peptide,[38] preferential cleavage may jeopardize unambiguous protein identification. However, the high-quality tandem mass spectra obtained from the QTOF spectrometer still provided sufficient sequence information to identify U6 snRNA-associated Sm-like protein LSm7 and TCTP.

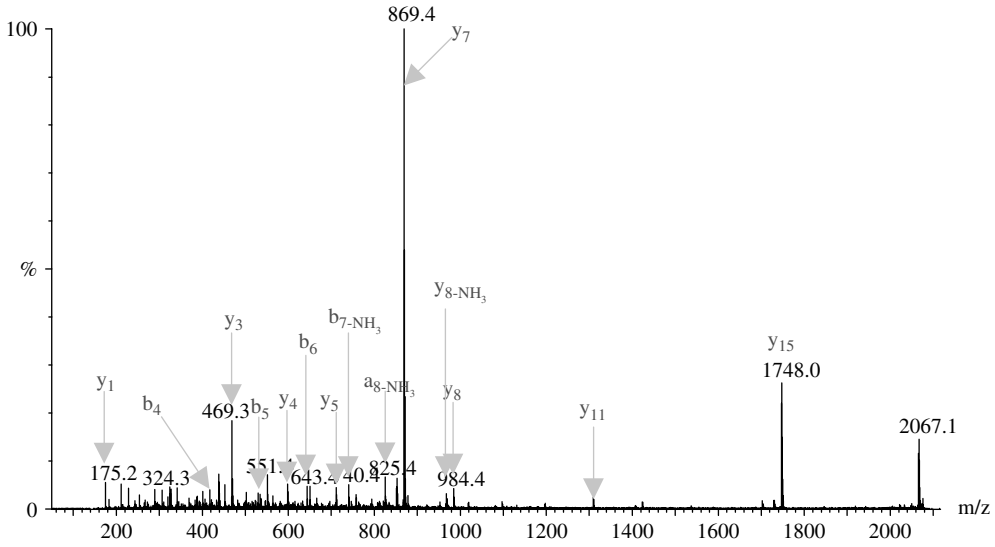

Figure 9. MALDI-QTOF tandem mass spectrum for 2066 (GFDPLLNLVLDGTIEYMR) from U6 snRNA-associated Sm-like protein LSm7.

CONCLUSION

Low-M_r proteins were separated from breast cancer cells using 2-D liquid separations. When the PMF method was used for identification, four proteins with $M_r < 20$ kDa were unambiguously identified. The identification was supported by an accurate M_r value obtained from ESI-TOFMS. However, other proteins remain unidentified owing to an insufficient number of proteins detected by MALDI-TOFMS. An enhancement of PMF can be achieved by including sequence information obtained from on-target light-induced methionine oxidation, but many proteins still remain unidentified. Significant improvements in protein identification were obtained when MALDI-QTOFMS was applied for analyzing the proteins with an insufficient number of peptides for PMF. The method achieves protein identification for proteins, which were not identified in PMF owing to internal fragmentation, few arginines in the sequence and incomplete tryptic digestion. In some cases, high-quality MS/MS data obtained from MALDI-QTOFMS overcome preferential cleavages and result in unambiguous protein identification when combined together with the M_r value of the intact protein. The accurate M_r confirms the protein identification achieved by using the MS/MS and in many cases the M_r information is necessary for correct protein identification, especially for modified forms.

Acknowledgments

We thank Dr Phillip Andrews of the University of Michigan Medical Center Proteomics Facility for use of the Q-TOF Ultima. This work was supported in part by the National Institutes of Health under grant R01 GM 49 500 (D.M.L.) and the National Cancer Institute under grants R21CA83808 (D.M.L., F.R.M.) and R01CA90503 (F.R.M., D.M.L.). Support was also generously provided by Eprogen, Inc. The MALDI-TOFMS instrument used in this work was funded by the National Science Foundation under grant DBI 99 874.

REFERENCES

1. Henzel WJ, Billeci TM, Stults JT. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 1993; **90**: 5011.
2. Fenyo D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* 2000; **11**: 391.
3. Yates JR III, Speicher S, Griffin PR, Hunkapiller T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 1993; **214**: 397.
4. James P, Quadroni M, Carafoli E, Gonnet G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 1993; **195**: 58.
5. Pappin DJC, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 1993; **3**: 327.
6. Krause E, Wenschuh H, Jungblut PR. The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal. Chem.* 1999; **71**: 4160.
7. Harrison AG. The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrom. Rev.* 1997; **16**: 201.
8. Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 1999; **71**: 2871.
9. Bienvenut WV, Hoogland C, Greco A, Heller M, Gasteiger E, Appel RD, Diaz J-J, Sanchez J-C, Hochstrasser DF. Hydrogen/deuterium exchange for higher specificity of protein identification by peptide mass fingerprinting. *Rapid Commun. Mass Spectrom.* 2002; **16**: 616.
10. Zhu H, Hunter TC, Pan S, Yau PM, Bradbury EM, Chen X. Residue-specific mass signatures for the efficient detection of protein modifications by mass spectrometry. *Anal. Chem.* 2002; **74**: 1687.
11. Hunter TC, Yang L, Zhu H, Majidi V, Bradbury EM, Chen X. Peptide mass mapping constrained with stable isotope-tagged peptides for identification of protein mixtures. *Anal. Chem.* 2001; **73**: 4891.
12. Pratt JM, Robertson DHL, Gaskell SJ, Riba-Garcia I, Hubbard SJ, Sidhu K, Oliver SG, Butler P, Hayes A, Petty J, Beynon RJ. Stable isotope labelling *in vivo* as an aid to protein identification in peptide mass fingerprinting. *Proteomics* 2002; **2**: 157.
13. Wise MJ, Littlejohn TG. Peptide-mass fingerprinting and the ideal covering set for protein characterisation. *Electrophoresis* 1997; **18**: 1399.
14. Brancia FL, Butt A, Beynon RJ, Hubbard SJ, Gaskell SJ, Oliver SG. A combination of chemical derivatization and improved bioinformatic tools optimizes protein identification for proteomics. *Electrophoresis* 2001; **22**: 552.

15. Beardsley RL, Karty JA, Reilly JP. Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* 2000; **14**: 2147.
16. Santner SJ, Dawson PJ, Tait L, Soule HD, Eliason J, Mohamed AN, Wolman SR, Heppner GH, Miller FR. Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res. Treat.* 2001; **65**: 101.
17. Yan F, Subramanian B, Nakeff A, Barder TJ, Parus SJ, Lubman DM. A comparison of drug-treated and untreated HCT-116 human colon adenocarcinoma cells using a 2-D liquid separation mapping method based upon chromatofocusing PI fractionation. *Anal. Chem.* 2003; **75**: 2299.
18. Fountoulakis M, Takacz B, Langen H. Two-dimensional map of basic proteins of Haemophilus influenzae. *Electrophoresis* 1998; **19**: 1819.
19. Keller BO, Wang Z, Li L. Low-mass proteome analysis based on liquid chromatography fractionation, nanoliter protein concentration/digestion, and microspot matrix-assisted laser desorption ionization mass spectrometry. *J. Chromatogr. B* 2002; **782**: 317.
20. Zhu K, Kim JK, Yoo C, Miller FR, Lubman DM. High sequence coverage of proteins isolated from liquid separations of breast cancer cells using capillary electrophoresis–time-of-flight MS and MALDI-TOF MS mapping. *Anal. Chem.* 2003; **75**: 6209.
21. Holmes MR, Giddings MC. Prediction of posttranslational modifications using intact-protein mass spectrometric data. *Anal. Chem.* 2004; **76**: 276–282
22. Jensen ON, Podtelejnikov AV, Mann M. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* 1997; **69**: 4741.
23. Yates JR III, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* 1995; **67**: 1426.
24. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra, TD, Udseth HR. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002; **2**: 513.
25. Goodlett DR, Bruce JE, Anderson GA, Rist B, Pasa-Tolic L, Fiehn O, Smith RD, Aebersold R. Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. *Anal. Chem.* 2000; **72**: 1112.
26. Sechi S, Chait BT. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Anal. Chem.* 1998; **70**: 5150.
27. Chen X, Smith LM, Bradbury, EM. Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal. Chem.* 2000; **72**: 1134.
28. Ogorzalek Loo RR, Loo JA, Du P, Holler T. *In vivo* labeling: a glimpse of the dynamic proteome and additional constraints for protein identification. *J. Ame. Soc. Mass Spectrom.* 2002; **13**: 804.
29. Wilkins MR, Gasteiger E, Tonella L, Ou K, Tyler M, Sanchez J-C, Gooley AA, Walsh BJ, Bairoch A, Appel RD, Williams KL, Hochstrasser DF. Protein identification with N- and C-terminal sequence tags in proteome projects. *J Mol. Biol.* 1998; **278**: 599.
30. Brancia FL, Oliver SG, Gaskell SJ, Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Commun. Mass Spectrom.* 2000; **14**: 2070.
31. Yates JR III, McCormack AL, Eng J. Mining genomes with MS. *Anal. Chem.* 1996; **68**: 534A.
32. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Ame. Soc. Mass Spectrom.* 1994; **5**: 976.
33. Loboda AV, Krutchinsky AN, Bromirski M, Ens W, Standing KG. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. *Rapid Commun. Mass Spectrom.* 2000; **14**: 1047.
34. Verhaert P, Uttenweiler-Joseph S, De Vries M, Loboda A, Ens W, Standing KG. Matrix-assisted laser desorption/ionization quadrupole time-of-flight mass spectrometry: an elegant tool for peptidomics. *Proteomics* 2001; **1**: 118.
35. Wattenberg A, Organ AJ, Schneider K, Tyldesley R, Bordoli R, Bateman RH. Sequence dependent fragmentation of peptides generated by MALDI quadrupole time-of-flight (MALDI Q-TOF) mass spectrometry and its implications for protein identification. *J. Ame. Soc. Mass Spectrom.* 2002; **13**: 772.
36. Tsaprailis G, Nair H, Somogyi A, Wysocki VH, Zhong W, Futrell JH, Summerfield SG, Gaskell SJ. Influence of secondary structure on the fragmentation of protonated peptides. *J. Ame. Chem. Soc.* 1999; **121**: 5142.
37. Gu C, Tsaprailis G, Breci L, Wysocki VH. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal. Chem.* 2000; **72**: 5804.
38. Qin J, Fenyo D, Zhao Y, Hall WW, Chao DM, Wilson CJ, Young RA, Chait BT. A strategy for rapid, high-confidence protein identification. *Anal. Chem.* 1997; **69**: 3995.