

Peptide mass fingerprinting

Bernd Thiede^a, Wolfgang Höhenwarter^b, Alexander Krah^{a,b}, Jens Mattow^c,
Monika Schmid^b, Frank Schmidt^{b,c}, Peter R. Jungblut^{b,*}

^a Department of Molecular Biology, Max Planck Institute for Infection Biology, Schumannstrasse 21/22, D-10117 Berlin, Germany

^b Core Facility Protein Analysis, Max Planck Institute for Infection Biology, Schumannstrasse 21/22, D-10117 Berlin, Germany

^c Department of Immunology, Max Planck Institute for Infection Biology, Schumannstrasse 21/22, D-10117 Berlin, Germany

Accepted 25 August 2004

Available online 12 January 2005

Abstract

Peptide mass fingerprinting by MALDI-MS and sequencing by tandem mass spectrometry have evolved into the major methods for identification of proteins following separation by two-dimensional gel electrophoresis, SDS-PAGE or liquid chromatography. One main technological goal of proteome analyses beside high sensitivity and automation was the comprehensive analysis of proteins. Therefore, the protein species level with the essential information on co- and post-translational modifications must be achieved. The power of peptide mass fingerprinting for protein identification was described here, as exemplified by the identification of protein species with high molecular masses (spectrin α and β), low molecular masses (elongation factor EF-TU fragments), splice variants (α A crystallin), aggregates with disulfide bridges (alkylhydroperoxide reductase), and phosphorylated proteins (heat shock protein 27). Helpful tools for these analyses were the use of the minimal protein identifier concept and the software program MS-Screener to remove mass peaks assignable to contaminants and neighbor spots.

© 2004 Elsevier Inc. All rights reserved.

Keywords: MALDI-MS; Minimal protein identifier; MS-Screener; Peptide mass fingerprinting; Proteome; Proteomics

1. Introduction

Protein identification by mass spectrometry (MS)¹ can be performed using sequence-specific peptide fragmentation or peptide mass fingerprinting (PMF), also known as peptide mass mapping [1]. The standard approach to identify proteins includes separation of pro-

teins by gel electrophoresis or liquid chromatography. Subsequently, the proteins are cleaved with sequence-specific endoproteases, most notably trypsin. Following digestion, the generated peptides are investigated by determination of molecular masses or generation of peptide fragments. For protein identification, the experimentally obtained masses are compared with the theoretical peptide masses of proteins stored in databases by means of mass search programs (Fig. 1) [2].

Matrix-assisted laser/desorption ionization mass spectrometry (MALDI-MS) is the most commonly used technique to perform PMF [3,4]. MALDI-MS is fast, robust, easy to perform, sensitive (low fmol range), accurate (low ppm range), tolerant to a certain level of various contaminants, and can be automated. The predominant detection of singly charged peptide molecules by MALDI-MS facilitates the evaluation of PMFs

* Corresponding author. Fax +49 30 28460507.

E-mail address: jungblut@mpiib-berlin.mpg.de (P.R. Jungblut).

¹ Abbreviations used: 2-DE, two-dimensional gel electrophoresis; CHCA, α -cyano-4-hydroxycinnamic acid; CID, collision-induced dissociation; DHB, 2,5-dihydroxybenzoic acid; ESI, electrospray ionization; LC, liquid chromatography; MALDI, matrix-assisted laser desorption/ionization; MM, number of mass values matched; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PMF, peptide mass fingerprinting; PSD, post-source decay; SC, sequence coverage; TFA, trifluoroacetic acid.

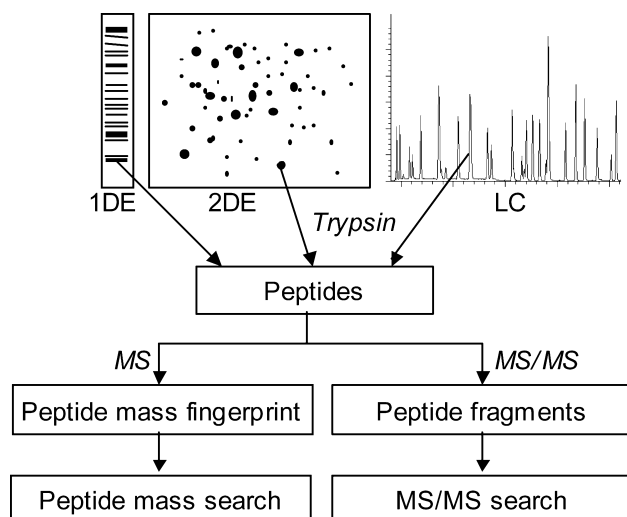


Fig. 1. Common procedures to identify proteins by mass spectrometry.

significantly. Although PMF is an effective tool for the identification of relatively pure proteins, it often fails to identify protein mixtures. Separation of complex protein samples by high-resolution two-dimensional gel electrophoresis (2-DE) is well adapted to protein identification with PMF. On the other hand, the application of PMF in combination with one-dimensional gel electrophoresis or liquid chromatography must be adjusted to the separation capacity.

A frequently used strategy to identify proteins by MS is to first generate a PMF because of the simplicity of the method. Sequence-specific peptide fragmentation is necessary if no protein was identified unambiguously. Peptide fragments can be generated directly by MALDI-MS using post-source decay (PSD)-MALDI-MS or collision-induced dissociation (CID)-MALDI-MS/MS. The relative new MALDI-ion traps, MALDI-Q-TOF and MALDI-TOF/TOF instruments are particularly suitable for this purpose. Alternatively, nanoelectrospray ionization tandem mass spectrometry (nano-ESI-MS/MS) is used frequently to achieve sequence information.

Although PMF is relatively simple and a standard procedure can be described, various factors influence the outcome of the analysis. Therefore, we would like to discuss these factors in more detail with emphasis on examples of protein identification by PMF from gel bands and 2-DE gel spots. In particular, the sample origin, sample purification, sample preparation for MALDI-MS, and factors influencing the mass search are discussed. Furthermore, the power and limitations of PMF are illustrated by the identification of phosphorylation sites, splicing variants, proteins with disulfide bridges, large proteins, and protein fragments. Another focus is the MS-Screener software which can be used for the advanced evaluation of PMFs.

2. Methods

2.1. Sample origin

Two-dimensional gel electrophoresis is the preferred method for protein separation prior to PMF. Gel separated proteins have to be enzymatically or chemically cleaved prior to mass analysis. In-gel digestion is used predominantly, but on-membrane digestions can be performed as well [5,6]. However, the blotting efficiency is dependent on the blotting conditions and different for high and low molecular proteins. Preparative 2-DE gels prior to protein identification by mass spectrometry are typically stained either by silver (without glutaraldehyde) or Coomassie Blue G-250. Silver staining even without glutaraldehyde reduces the sequence coverage considerably [7]. Fluorescent labeling of proteins can be achieved with ruthenium II tris (bathophenanthroline disulfonate) or SyproRuby, but needs special equipment [8]. Different laboratories favor distinct staining techniques [9]. As a rule of thumb, spots detectable by Coomassie Blue G-250 can be identified by PMF.

Trypsin is the favored enzyme for PMF. Trypsin is relatively cheap, highly effective, and generates peptides with an average size of about 8–10 amino acids, ideally suited for analysis by MS. The endoprotease Lys-C is sometimes used to receive a higher sequence coverage in comparison to trypsin because longer peptides are generated. The endoprotease Glu-C (V8) cleaves proteins C-terminal of glutamic acid or aspartic acid/glutamic acid dependent on the buffer conditions. However, Glu-C is not commonly used, most likely because of the formation of many autoproteolytic peptides. Asp-N and Arg-C are relative expensive and only employed for special purposes. Cyanogen bromide is an effective but toxic chemical which cleaves proteins C-terminally to methionines by formation of homoserine lactone. Large peptides are usually generated due to the low number of methionines per protein. For some applications this maybe beneficial, although it is not favorable for mass spectrometry. However, a subsequent cleavage with trypsin can be performed to overcome this drawback.

2.2. Sample purification

The generated peptide mixture must be concentrated and desalted prior to mass analysis if no accurate crystallization was obtained on the MALDI target or if the generated mass spectrum was of insufficient quality for protein identification.

Desalting and concentration of peptide mixtures can be performed on reversed-phase microcolumns. For this purpose, either commercially available prepacked reversed-phase columns (e.g., ZipTips, see Section 3.2) or self-made packed GELoader tips can be used. The manual fabrication of GELoader tips filled with Poros R2 or

other beads requires some experience. Sequential application of, e.g., Poros R2, Oligo R3, and graphite powder increases the sequence coverage [10]. StageTips simplify the production of self-made columns using reversed-phase beads incorporated in poly(tetrafluoroethylene) membrane disks [11].

The signal-to-noise ratio is slightly improved after purification of the peptides by reversed-phase microcolumns. Loss of peptides is usually not observed. Perhaps one may lose some small hydrophilic peptides due to the use of ZipTips which are not essential for protein identification. Stepwise elution (e.g., by subsequent addition of 10, 20, and 50% acetonitrile) leads to a slight separation, which may help in special applications. However, three spectra must be generated and combined for the identification.

2.3. Sample preparation

The dried droplet technique is the predominantly applied MALDI-MS sample preparation technique. This method is robust, simple, and effective. Alternatively, the thin layer technique is preferred by some groups, but seems to be more delicate to the material of the sample holder, acid, and solvent [12]. Favored matrices for MALDI-MS are 2,5-dihydroxybenzoic acid (DHB) and α -cyano-4-hydroxycinnamic acid (CHCA). The combination of both matrices yielded slightly improved performance [13]. Minimization of the sample size increases the sensitivity. Small volumes should be used for standard metal plates. On the other hand, hydrophilic sample anchors are efficient for the generation of small spots [14].

2.4. Mass search

After recording a mass spectrum, the monoisotopic peaks must be labeled. Peak detection and labeling is performed by mass spectra analysis programs, but usually require manual inspection and interactive correction. A correct isotopic distribution and a signal-to-noise >2 are minimal requirements for the definition of mass peaks. The determination of the first isotope with accuracy is sometimes difficult for masses with a mass >2500 Da, depending on the resolution of the peak. The mass accuracy of the measurements strongly depends on the mass spectrometer. Improving the mass accuracy by internal calibration is one way to reduce false-positive protein matches. For internal calibration, peptides with known amino acid sequence can be added to the sample or frequently observed contaminant masses can be used, e.g., the autoproteolytic tryptic peptides such as 842.51 and 2211.10 Da in their protonated form.

Mascot [15], MS-Fit [16], and Profound [17] are the most frequently used internet-accessible search programs for PMF. In principle, the same result will be

obtained with all programs using the same parameters. A prerequisite for successful PMF is the existence of the protein in the database. However, the sequencing of many genomes within the last decade has significantly increased the applicability of PMF. As an example, PMF by means of Mascot will be described in more detail (see Section 3.4).

3. Description of procedures

3.1. Manual trypsin digest of proteins

During gel electrophoresis, gel staining, and sample preparation for mass analysis, special caution must be taken to avoid contamination of the protein samples with keratin. For this purpose, gels should be sealed as soon as possible after staining. It is important to wear gloves and a laboratory coat. Gloves must be rinsed with distilled water. Furthermore, all solutions used for the enzymatic digestion should be stored in small aliquots and used only once. Tubes should be inspected for dust particles. The quality of the tubes must be checked to avoid contamination with polymers. Sequencing or LC grade chemicals and enzymes should be used exclusively.

Wash empty 500 μ l tube twice with 200 μ l acetonitrile. Add 70 μ l of 200 mM ammonium hydrogen carbonate (NH_4HCO_3), pH 7.8, per tube. For spot excision, take an eppendorf combitip (500 μ l) and cut it with a scalpel at about half height. Then, pull the plunger up, put the tip on the spot, turn it slightly, and transfer the spot into the tube by pressing the plunger out. This procedure leads to a uniform spot size. In addition, it is easier to use these tips than a scalpel. Cut the spots with a pair of sharp tweezers to increase the gel surface. Wash the gel material by shaking the tube 30 min at 37 $^\circ\text{C}$. Discard the washing solution. Shrink the gel material by adding 70 μ l of 200 mM NH_4HCO_3 , pH 7.8/acetonitrile (4:6), shake 30 min at 37 $^\circ\text{C}$. Remove the shrinking solution. Rehydrate the gel material by addition of 70 μ l of 50 mM NH_4HCO_3 , pH 7.8, shake 30 min at 37 $^\circ\text{C}$. Discard the rehydration solution. Shrink the gel material by adding 70 μ l acetonitrile and incubate for 2 min. Again, remove the shrinking solution. Add 0.1 μ g trypsin in 25 μ l of 25 mM NH_4HCO_3 , pH 7.8. Use a stock solution of 20 μ g trypsin reconstituted in 100 μ l of 50 mM acetic acid. This is stored in 10 μ l aliquots at -20°C . For protein digestion, mix one aliquot (10 μ l = 2 μ g) with 500 μ l of 50 mM NH_4HCO_3 and use 25 μ l (0.1 μ g trypsin) of the resulting solution per sample. The protein digest is performed at 37 $^\circ\text{C}$ for 4–16 h. Add 10 μ l of 10% trifluoroacetic acid (TFA) to stop the digestion. Collect supernatant in 200 μ l tube. Add 20 μ l 0.3% TFA/60% acetonitrile to elute the remaining peptides out of the gel material. Collect supernatant again and pool the corresponding supernatants in a 200 μ l tube. Reduce volume by means

of a Speedvac. The evaporation is a critical point of the procedure which may cause loss of sample. However, we used a Speedvac for many years to concentrate the samples prior to mass analysis. Alternatively, 0.5 µl of the supernatant after the digestion with trypsin mixed with 0.5 µl of 10% formic acid and 1 µl of matrix can be used directly with anchor chips.

3.2. Purification of peptides by ZipTip_{µ-C18}

Wash empty tube twice with 100 µl elution solution (1% formic acid/acetonitrile (1:1)). Add 10 µl elution solution (1% formic acid/acetonitrile (1:1)) into empty tubes. Dissolve sample in 10 µl of 1% formic acid. Wash ZipTip twice with 15 µl (1% formic acid/acetonitrile (1:2)). Equilibrate ZipTip twice with 15 µl of 1% formic acid. Load sample onto ZipTip (three times up and down). Wash peptides bound onto ZipTip three times with 15 µl water. Elute sample into elution solution (three times up and down) into a tube and concentrate sample prior to MALDI-MS analysis. The same ZipTip can be used repeatable without cross contamination if the procedure is performed carefully.

3.3. Dried droplet sample preparation

Dissolve peptides (as described in Section 3.1) by adding 1 µl of 1% formic acid. The remaining solution can be directly analyzed by LC-ESI-MS/MS. Be cautious to dissolve the peptides in a very small volume. Make sure that you wash the walls of the tube repeatedly with the small droplet. Mix 0.3 µl matrix and 0.3 µl sample solution and transfer the resulting solution to the sample plate. Work as quickly as possible to avoid evaporation of the solvent. The matrix solution of choice depends on several factors, in particular on the material of the sample holder. Approximately, use as matrix DHB (5 g/L) in 0.3% TFA:acetonitrile (2:1) for anchor chips, and DHB (50 g/L), or CHCA (20 g/L) in 0.3% TFA:acetonitrile (2:1) for metal plates. The internal standard peptides can be premixed with the matrix. Peptide mass peaks occurring due to autolysis of trypsin (porcine) such as 842.51 and 2211.10 Da can also be used for internal calibration.

3.4. Peptide mass fingerprinting with Mascot

Peak detection must be performed after calibration of the mass spectrum with internal standards. Autoproteolytic peptide masses of trypsin as well as peptide added for internal calibration are omitted for database searches. The list of peptide masses can be transferred into the peptide mass fingerprint search program Mascot (http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF) as data file or by copy/paste of the mass list into query. The NCBI database is relatively comprehensive, whereas the Swiss-

Prot database is less complete, but seems to be annotated with high accuracy. Furthermore, detailed information about the proteins is included in Swiss-Prot entries and several tools can be applied to the sequence analysis of the proteins directly. The experimentally determined peptide masses may be compared against the entire NCBI database (taxonomy: all entries). However, if the organism is specified, the mass search time can be reduced significantly. Keratin contaminating masses must be considered, if another organism than human is used for the database search. The enzyme used for proteolytic cleavage needs to be specified. Allow up to 1 missed cleavage is recommended, because a higher number significantly increases the number of theoretical peptide masses. Peptides with one missed cleavage site occur frequently whereas most of the observed peptides possess <2 missed cleavages. Carbamidomethyl has to be considered as fixed modification if iodoacetamide was used to modify cysteines. Acetyl (Protein) for eukaryotic proteins, oxidation (M), and pyro-Glu (N-term Q) are common variable modifications. Propionamide (C) must be considered if the samples were not treated with iodoacetamide. The peptide tol. ±, the error window on experimental peptide mass values, is dependent on the mass spectrometer used and the accuracy of the calibration.

The result of a peptide mass search with Mascot contains a lot of information. First, the probability based mowse score is very important. The difference between random and significant protein scores should be as high as possible. As a rule of thumb employing all protein entries, a protein is identified with a score >100. A score within the highest random value (usually between 60 and 75) and 100 displays a good candidate protein. Second, the concise or full protein summary report should be considered. By clicking onto the accession number of the first hit more detailed protein information is displayed. The nominal mass and the pI value must be in accordance with the experimental data from gel electrophoresis. If this is not the case, protein fragments or adducts should be considered. Furthermore, the sequence coverage (SC) and the number of mass values matched (MM) are very important. Here, we revised the reliability of the identification again by the multiplication of SC [%] and MM. A protein is considered to be identified by a value >300. The difference between the number of mass values searched and the number of mass values matched should be as small as possible. If the difference is high, the same sample may contain additional protein components. The average error value must be in the typical range of the experience with the mass spectrometer used. An inaccurate calibration is recognizable by the error of the single matched masses. Erratic fluctuations of the error values, in particular in the mass range of 1000–2000 Da, are inconvenient. After sorting the peptides by increasing mass, the matched and observed masses should be

compared. The most intense peaks should be assignable to the identified protein. In general, arginine-containing peptides in MALDI-MS derived mass fingerprints after tryptic digestion give rise to the most intense peaks [18]. The “search unmatched” function at the concise or full protein summary report is a useful tool to reevaluate unassigned masses.

Unassigned mass peaks may occur due to tryptophan oxidation [19], methylation of aspartic acid, and glutamic acid-rich peptides because of the staining procedure in acetic acid/methanol [20]. In addition, it is conceivable that the analyzed peptides contain a higher number of missed cleavage sites than considered for the search [19]. Further unusual modifications may arise during gel electrophoresis [21]. A comprehensive overview of observing a particular peptide based on a combination of chemical considerations with estimated probability for PMF was accomplished [22]. As a note, carbamylation and deamidation due to sample preparation were condemned as untrustworthy [23].

3.5. Special applications of PMF

3.5.1. Phosphorylated proteins

Phosphorylation/dephosphorylation of serine, threonine, and tyrosine residues modulates the function of proteins and is one of the most common cellular regulatory mechanisms. To detect phosphorylated peptides PMFs of individual enzymatic digests are often comparatively analyzed before and after alkaline phosphatase treatment [24]. A single hydrogenphosphate adduction leads to an 80 Da shift in molecular mass. Consequently, in case of a single phosphorylation event an 80 Da increment should be detectable within mass peaks corresponding to the unmodified and modified peptide in a MALDI-MS experiment. Treatment of the sample with alkaline phosphatase catalyzes the hydrolysis of the phosphate bond quenching the $M_r + 80$ Da of the two mass peaks in a subsequent experiment thus demonstrating the phosphorylation of the corresponding peptide. For more complex peptide mixtures enrichment of phosphorylated peptides using immobilized metal affinity chromatography (IMAC), e.g., with ZipTip_{MC} is recommended prior to enzymatic dephosphorylation.

Mammalian small heat shock proteins are known to undergo extensive phosphorylation as an answer to cellular stress such as a radical rise in temperature or the advent of other adverse environmental factors [25,26]. Phosphorylation sites on the murine small heat shock protein HSP 25/27 were identified *in vivo* and *in vitro* at serine residues 15 and 86 [27]. A PMF generated from a tryptic digest of a 2-DE spot was conclusively identified as murine HSP 25/27 by the Mascot search algorithm. Closer investigation revealed two peaks with the masses 1005.77 and 1085.76 displaying the 80 Da mass shift characteristic of a phosphorylation (Fig. 2). The mass

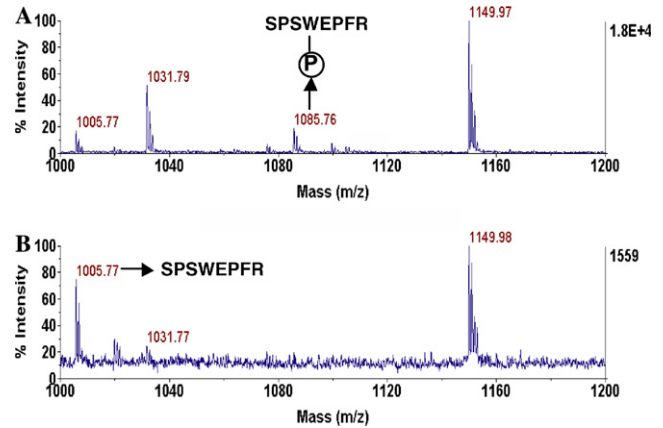


Fig. 2. Detection of phosphorylated peptides from 2-DE separated murine HSP 25/27 using alkaline phosphatase and MALDI-MS. PMF region of murine HSP 25/27 before alkaline phosphatase treatment shows a mass peak at 1085.76 Da indicative of a phosphorylated peptide (A). The same mass peak was quenched after alkaline phosphatase treatment and the mass peak corresponding to the dephosphorylated peptide at 1005.77 Da increased in relative intensity (B). The peptide sequence corresponding to the mass peaks is shown with the known phosphorylation site.

peak 1005.77 was assigned to amino acid positions 13–20 on the HSP25/27 sequence, the mass peak 1085.76 remained unassigned. These indications made a phosphorylation at serine 15 seem likely. Five microliters of 50 mmol NH_4HCO_3 , 5% acetonitrile buffer was added to 0.5 μl of the sample. One microliter of shrimp alkaline phosphatase was also added. This sample approach was incubated for 1 h at 37 °C, lyophilized, and resolved in 2.5 μl of 33% acetonitrile, 0.1% TFA. Half a microliter was spotted onto a MALDI target and analyzed. The greater mass peak 1085.76 was quenched while the lesser of the two remained (Fig. 2), indeed confirming the previously reported phosphorylation at serine residue 15.

3.5.2. Splice variants

Pre-mRNA splicing, the removal of introns from mRNA precursors, is an important mechanism for regulating gene expression in higher eukaryotes. A substantial proportion of higher eukaryotic genes produce multiple proteins in this way from a single transcript. In many cases, the alternatively spliced exon encodes a protein domain that is functionally important for catalytic activity or binding interactions, the resulting proteins can exhibit different or even antagonistic activities [28].

α crystallins constitute about 30% of total water soluble protein in the adult lens and are thus the major abundant lenticular protein class. They form high molecular mass multimers *in vivo*. These aggregates are made up of the distinct α crystallin subclasses α A and α B and have a molecular mass of about 800 kDa. In addition to themselves being a fundamental composite of molecular lens structure, α crystallins, much like the related small heat shock proteins, function as molecular chaperones

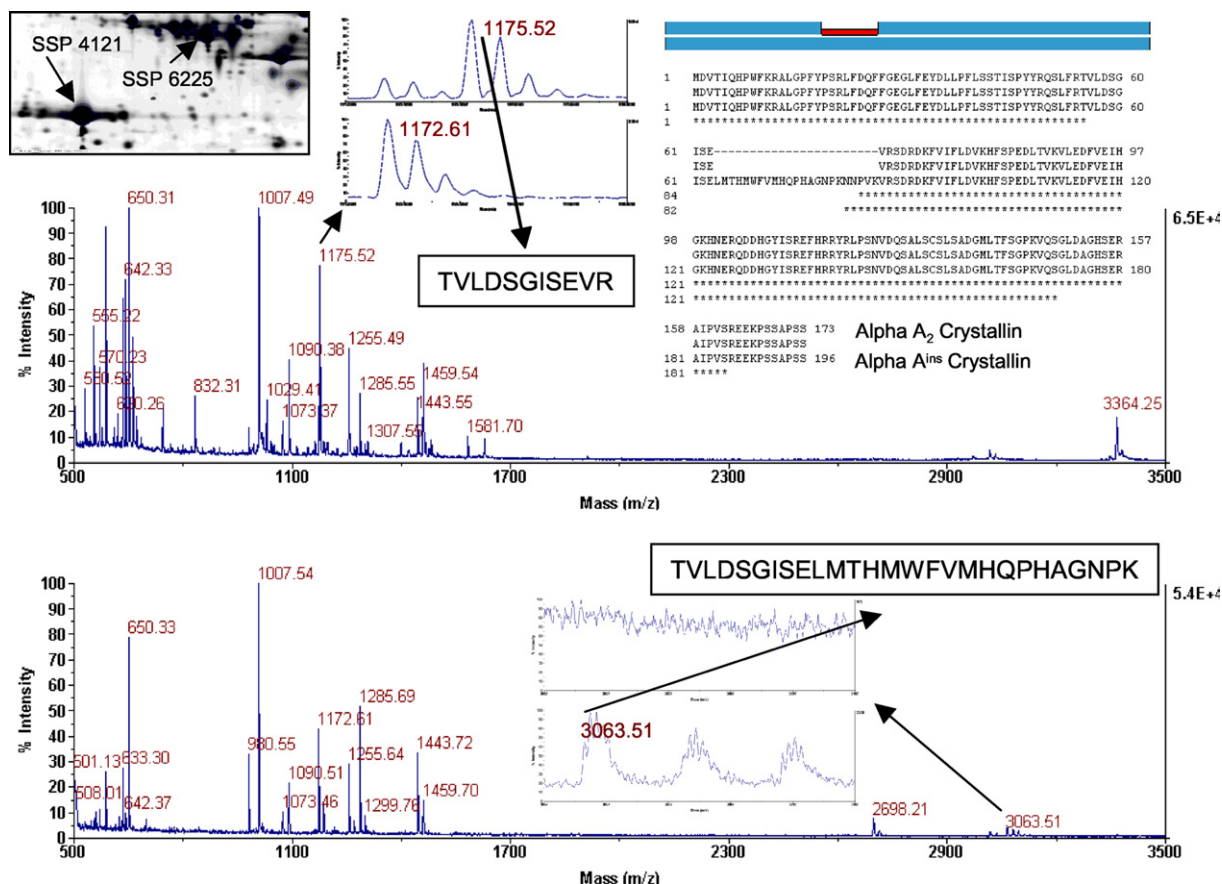
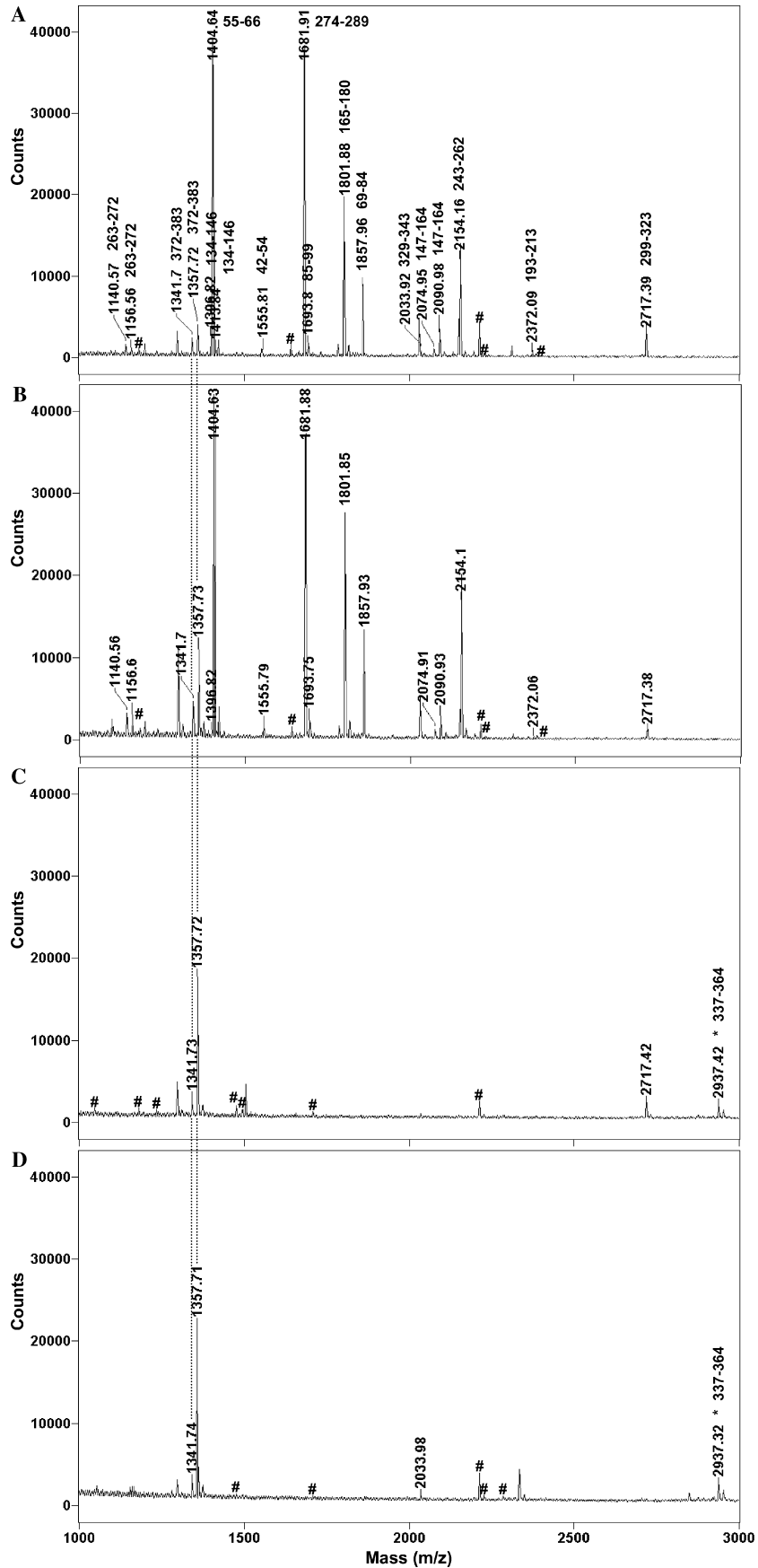


Fig. 3. Identification of alternative splice variants of crystallin A separated by 2-DE and analyzed by PMF. Analyzed spots containing the two splice variant α A crystallins and respective MALDI-MS PMFs of the tryptic digest are shown. Mass peaks assigned to peptides, respectively, bridging or filling the intron gap created by alternative splicing are cut away. Peptide sequences are shown. Both polypeptide sequences were aligned using the TBLASTN engine with the inserted peptide indicated.

helping to preserve lenticular molecular order and transparency by keeping proteins inclined to aggregate and precipitate due to abnormal modification in solution [29]. α A crystallin is a highly conserved polypeptide of 173 amino acid residues identical in mouse and rat. An alternative α A crystallin polypeptide, α A^{ins} crystallin, comprising 196 amino acid residues is found in lenses of the rodent families Muridae (mouse, rat) and Cricetinae (hamster, gerbil) [30]. This polypeptides primary structure matches the shorter α A, termed α A₂ crystallin primary structure save for an additional 23 amino acids

inserted between positions 63 and 64 of the shorter sequence. The shorter α A₂ mRNA is 5–10 times more abundant than the longer α A^{ins} mRNA, the proteins are thus also termed α A crystallin major and minor component, respectively. We separated the 10-day-old lenticular proteome of mouse strain C57BL by 2-DE and analyzed selected spots by MALDI-MS. α A₂ and α A^{ins} crystallins clearly migrated to gel positions consistent with their theoretical M_r and pI and were identified by PMF. The PMF generated from the spot containing α A₂ crystallin (SSP 4121) contained a monoisotopic

Fig. 4. Identification of truncated variants of elongation factor EF-TU (tuf) by combination of 2-DE, PMF, and MPI. MALDI-MS spectra of four 2-DE spots identified as tuf of *M. tuberculosis* H37Rv are shown. Two of these spots, whose spectra are depicted in (A and B), displayed an electrophoretic mobility consistent with the theoretical M_r (44.4 kDa) and pI (5.2) of full-length tuf. These spots were unambiguously identified by conventional PMF. In the case of (A) 18 tuf-specific masses were detected. Numbers given in brackets specify the AA residues (start and end) of the corresponding peptides. Peaks marked with # are putative contaminant masses, e.g., masses assignable to keratins. Unassigned and putative contaminant masses were deleted to generate an adjusted template spectrum. Using the program MS-Screener this spectrum was compared with 446 spectra from 430 distinct 2-DE spots of *M. tuberculosis* H37Rv CSN. The MS-Screener analysis revealed eight previously unidentified spots whose spectra contained ≥ 3 tuf-specific peptide masses. Examples are shown in (C and D). These spots showed an electrophoretic mobility (M_r , 10–13 kDa; pI , 7–9) markedly differing from the expected one of full-length tuf. However, the targeted ESI-MS/MS analysis of these eight spots confirmed the proposed identity. Note, that only tuf-specific peptide mass peaks are indicated. Mass peaks marked with * are tuf-specific, but were not detected in the template spectrum (A). Dashed vertical lines indicate tuf-specific peptide masses observed in all four spectra.



peak with 1175.52 Da which was absent in the PMF generated from the spot containing the splice variant. The amino acid sequence assigned to this mass by the Mascot search algorithm corresponds to positions 55–65 on the αA_2 sequence bridging the intron gap. Conversely, the PMF generated from the αA^{ins} containing spot (SSP 6225) displayed the monoisotopic peak at 3063.45 Da, corresponding to positions 55–81 on the αA^{ins} sequence, filling the intron gap with the insert peptide encoded by the alternative exon. Alternative splice variants are thus conclusively shown on the proteins themselves by PMF (Fig. 3).

3.5.3. Proteins with disulfide bridges

One protein may appear within one gel in different bands (1-DE) or spots (2-DE) due to modifications. One example for a modification is the disulfide bridge formation between cysteine residues. On one hand, this may occur within a single protein changing its tertiary structure and electrophoretic mobility. On the other hand, different molecules can be cross-linked by disulfide bridges leading to the formation of multimers. Different multimers of a protein are found on a gel at different molecular masses. The formation of disulfide bridges may occur *in vivo* as well as during sample preparation. Usually, reducing conditions caused by the use of DTT or mercaptoethanol prevent formation of disulfide bridges. Nevertheless, reduction is not always complete, especially if the concentration of the reducing agent decreases during gel electrophoresis. However, this may be prevented by alkylation of thiol-groups prior to separation, e.g., by iodoacetamide.

The existence of multimers may be indicated if several bands or spots contain the same protein with similar sequence coverages. In this case, it is worthwhile to examine these bands or spots in more detail. It is important to consider several possible modifications of cysteines during database searches. Non-polymerized acrylamide, which remains in low concentrations in the gel even after polymerization, is known to alkylate thiol-groups to form propionamide (mass difference = 71 Da). Other chemicals used for 2-DE can also lead to modifications [31]. If cysteines are involved in disulfide bridges, the respective peptides cannot be matched to the theoretically digested protein of the database. Cysteine residues involved in disulfide bridges can be found by comparison of the PMF coverage in such bands or spots. In beneficial cases, if one cysteine is missing in each spectrum, it is conceivable that the sites of linkage were found. An example for the discovery of a protein and its dimer within the same 2-DE gel was described in a study of cellular proteins of *Helicobacter pylori* [32]. The protein alkyl hydroperoxide reductase was identified in two different spots of which one displayed about the double theoretical molecular mass. Therefore, it was assumed that this spot represents a dimer of this protein. Two cysteine

containing peptides, as confirmed by MALDI-MS/MS, were found in the mass spectrum of the main spot but not detected in the mass spectrum of the dimer spot. Either one or both of them will supposedly be involved in the formation of disulfide bridges to create the dimer.

3.5.4. Low molecular mass proteins and protein fragments

PMF by MALDI-MS often fails to identify low molecular mass proteins and protein fragments due to the small number of detectable peptides. To overcome these limitations, PMF may be complemented by sequence generating MS/MS methods. However, it has also been suggested that reliable identification of proteins upon proteolytic cleavage is possible by a minimal set of experimentally derived peptide masses generated by MALDI-MS, termed a minimal protein identifier (MPI) [33]. The MPI approach is based on the notion that proteins yield characteristic peptide masses upon enzymatical cleavage, which may serve as signature masses. Using the MPI approach MALDI-MS spectra of proteins to be identified are compared to those of previously identified counterparts. In analogy to the ChemScore [22], which is calculated on the basis of chemical properties, the MPI approach considers a peptides detection probability by MALDI-MS. In contrast, conventional MALDI-MS PMF is based on correlating experimental and theoretical mass data stored in protein sequence databases without considering which peptides are efficiently generated by proteolytic protein cleavage and likely to be detected by MALDI-MS.

The MPI approach has recently been applied to systematically track the distribution of proteins in high-resolution 2-DE patterns of *Mycobacterium tuberculosis* H37Rv culture supernatant (CSN) proteins. The dataset investigated comprised all 446 MALDI-MS spectra recorded to establish the subproteome of *M. tuberculosis* H37Rv CSN [34]. The dataset spectra were compared with spectra of selected CSN proteins previously identified by MALDI-MS PMF, e.g., of mycobacterial elongation factor EF-TU (*tuf*) by the program MS-Screener [35]. This analysis revealed MALDI-MS spectra of 22 distinct spots containing ≥ 3 *tuf*-specific peptide masses, eight of which had previously not been identified as *tuf* by PMF. The targeted ESI-MS/MS analysis of these spots confirmed the proposed identity. The newly identified *tuf* protein species displayed an apparent M_r between 10 and 13 kDa and a *pI* ranging from 7.0 to 9.0, markedly differing from the theoretical M_r and *pI* values of full-length *tuf* (M_r : 44.4 kDa; *pI*: 5.2). The MALDI-MS spectra of four spots, two identified as full-length and two as truncated *tuf*, are shown in Fig. 4. The identification of the truncated *tuf* variants by PMF obviously failed due to their small size ($\sim 1/4$ of the full-length protein) resulting in a low number of matching peptides (≤ 4) and low sequence coverage ($\leq 17\%$). All assignable peptide masses of these protein species matched carboxy-terminal *tuf*-peptides [34].

3.5.5. High molecular mass proteins

The identification of large proteins by PMF raises problems, particularly at molecular masses higher than 100 kDa. Although many peptides are generated, the typical low sequence coverage by PMF prejudices the reliability of the result. Furthermore, only few protein spots are commonly detected by 2-DE in the high M_r range. Complementary, one-dimensional SDS-PAGE can be performed for high M_r proteins albeit with lower separation capability.

In the following, the difficulty to identify high M_r proteins is exemplified by a protein mixture of a human adenocarcinoma cell line (AGS). The AGS cells were infected by *H. pylori* and the cellular proteins were separated by the procedure of Laemmli [36]. The high- M_r range of the resulting SDS-PAGE gel (Fig. 5) shows several well-resolved bands. The band labeled with an arrow was digested in-gel by trypsin and the resulting peptides were analyzed by MALDI-MS with the direct measurement method [37]. The PMF comprised several hundred peaks and 140 were accepted as well resolved in the mass range between 500 and 3000 Da. The mass range between 1000 and 1500 Da (Fig. 5) exemplifies the large number of peptides only within this range. The database search with Mascot resulted in four proteins with significant probability score. Nevertheless, it was not possible to decide, if all these four proteins were present in the band or not. Because many of the peaks

were assignable with two, three or even all of the candidate proteins, it was unlikely that all proteins were present. In such cases, sequence information is necessary for an unambiguous identification. MALDI-MS/MS performed with a TOF/TOF instrument led to the identification of spectrin α and spectrin β . We could not find evidence for the presence of the other two candidate proteins predicted by the PMF. This example illustrates that it is advantageous to support PMF by sequence information for the unequivocal identification of proteins in the high molecular mass range.

3.5.6. Iterative analysis of PMFs

An approach for iterative data analysis and a software, MS-Screener (<http://www.mpiib-berlin.mpg.de/2D-PAGE/download.html>), were developed for in-depth analysis of PMFs [35]. In the first step of the procedure contaminant masses, i.e., masses matching matrix, keratins, autolysis products of trypsin or dye, are detected and deleted from the spectra using the program MS-Screener. Subsequently, peptide mass peaks caused by neighboring spots are determined by cluster analysis of the adjusted spectra and also deleted. The resulting spectra are analyzed by PMF using search algorithms such as Mascot. Following protein identification, the original spectra are systematically re-evaluated for the presence of masses matching the major components as well as contaminants. Mass peaks not yet assigned may be due

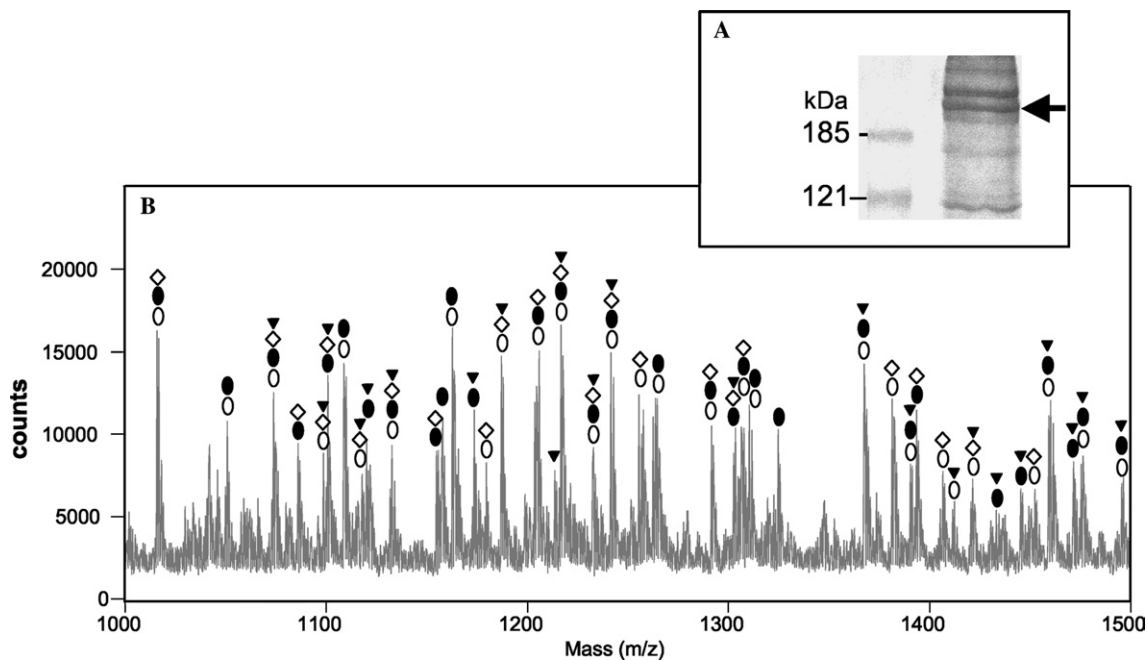


Fig. 5. PMF of a protein band from an SDS-PAGE gel with a high molecular mass >200 kDa. The SDS-PAGE gel of a cell preparation of AGS-cells is shown in (A), with the band analyzed labeled with an arrow. (B) MALDI-MS tryptic peptide mass fingerprint. Only the mass range 1000–1500 is shown comprising 43 labeled peaks. In the total mass spectrum more than 140 peaks were labeled. Filled triangle: trichohyalin, M_r 247 kDa, sequence coverage 25%, score 216. Empty diamond: CENP-E protein, M_r 311 kDa, sequence coverage 30%, score 223. Filled ellipse: spectrin α , M_r 284 kDa, sequence coverage 37%, score 330. Empty ellipse: spectrin β , M_r 274 kDa, sequence coverage 43%, score 422.

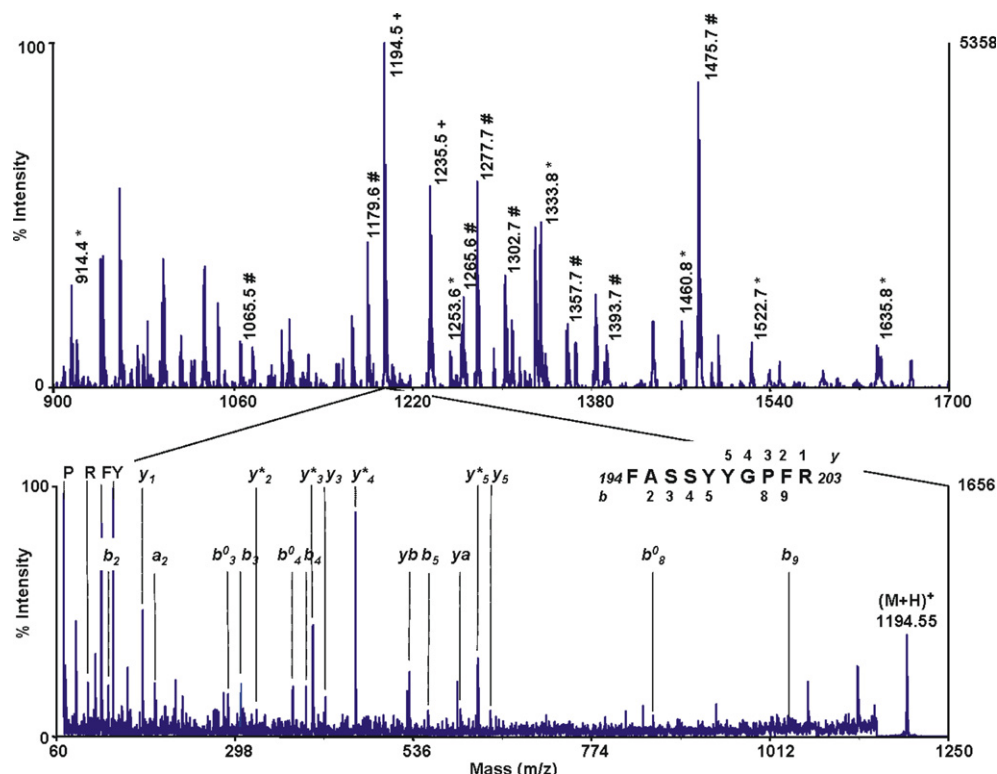


Fig. 6. Identification of two proteins within one 2-DE spot by iterative PMF data analysis and MS/MS. The MALDI-MS spectrum of a spot derived from a 2-DE gel of cytosolic proteins of *H. pylori* is shown. The analysis of the entire dataset by the program MS-Screener identified several contaminant masses marked with # that were eliminated prior to further analysis. The re-analysis of the adjusted spectrum revealed response regulator (HP0703) as the major spot component. Mass peaks marked with * are matching HP0703. A neighbor spot contamination of the spot by aminolevulinic acid dehydratase (HP0163) was predicted by cluster analysis and verified by MALDI-MS/MS (1194.5 and 1235.5 Da; marked with +). Note that all masses matching HP0703 display a relatively low intensity. In contrast, both masses matching the secondary component HP0163 show a relatively high intensity.

to post-translational modifications or additional protein components. Post-translational modifications can be detected using the FindMod tool (<http://us.expasy.org/tools/findmod/>) and validated by MS/MS. Additional protein components can be identified by re-evaluating the unassigned peaks by PMF and/or by MS/MS. This approach has been shown to be superior to conventional PMF [32,35] and facilitates in-depth analysis of the 2-DE distribution of proteins as well as of the protein composition of single 2-DE spots.

This iterative approach has recently been applied for comprehensive analysis of a dataset containing 480 MALDI-MS spectra derived by the analysis of distinct 2-DE spots representing cytosolic proteins of *H. pylori* [35]. For several spots, this analysis led to the identification of additional protein components. An example is shown in Fig. 6. The primary analysis of the PMF by Mascot did not reveal unequivocal protein identification. However, one protein component, response regulator (HP0703), was identified by application of the iterative approach and re-analysis of the adjusted data. In addition, a neighbor spot contamination by aminolevulinic acid dehydratase (HP0163) was predicted by cluster analysis of the dataset and validated by MALDI-MS/MS.

4. Concluding remarks

The development of MALDI-MS and ESI-MS and the rapid growth of databases have revolutionized the identification of proteins. PMF by MALDI-MS is the key approach for high throughput identification of well-separated proteins. Mass accuracy, sensitivity, and automatic measurement have been improved noticeably within the last decade for MALDI-TOF-MS, facilitating protein identification by PMF with high confidence. Furthermore, it has been recognized that 2-DE spots and their PMFs contain much more information than initially expected. A 2-DE spot can contain several protein components and a single protein can give rise to several 2-DE spots due to modifications. Protein species defined by differential chemical structure of one protein can occur as spot series but may also be widely distributed across a 2-DE pattern. The analysis of proteins at the protein species level requires high sequence coverage and in-depth evaluation of the spot composition. PMFs can already include information about the mode and site of protein modifications, which need to be confirmed by sequencing methods.

Acknowledgments

The authors acknowledge the Bundesministerium für Bildung und Forschung (BMBF 031U/107A-031U/207A and 0313029A) for financial support.

References

- [1] R. Aebersold, D.R. Goodlett, *Chem. Rev.* 101 (2001) 269–295.
- [2] R. Aebersold, M. Mann, *Nature* 422 (2003) 198–207.
- [3] K. Gevaert, J. Vandekerckhove, *Electrophoresis* 21 (2000) 1145–1154.
- [4] W.J. Henzel, C. Watanabe, J.T. Stults, *J. Am. Soc. Mass Spectrom.* 14 (2003) 931–942.
- [5] P.L. Courchesne, R. Luethy, S.D. Patterson, *Electrophoresis* 18 (1997) 369–381.
- [6] J. Fernandez, L. Andrews, S.M. Mische, *Anal. Biochem.* 218 (1994) 112–117.
- [7] C. Scheler, S. Lamer, Z. Pan, X.P. Li, J. Salnikow, P. Jungblut, *Electrophoresis* 19 (1998) 918–927.
- [8] T. Rabilloud, J.M. Strub, S. Luche, A. Van Dorsselaer, J. Lunardi, *Proteomics* 1 (2001) 699–704.
- [9] B. Moritz, H.E. Meyer, *Proteomics* 3 (2003) 2208–2220.
- [10] M.R. Larsen, S.J. Cordwell, P. Roepstorff, *Proteomics* 2 (2002) 1277–1287.
- [11] J. Rappsilber, Y. Ishihama, M. Mann, *Anal. Chem.* 75 (2003) 663–670.
- [12] J. Gobom, M. Schuerenberg, M. Mueller, D. Theiss, H. Lehrach, E. Nordhoff, *Anal. Chem.* 73 (2001) 434–438.
- [13] S. Laugesen, P. Roepstorff, *J. Am. Soc. Mass Spectrom.* 14 (2003) 992–1002.
- [14] M. Schuerenberg, C. Luebbert, H. Eickhoff, M. Kalkum, H. Lehrach, E. Nordhoff, *Anal. Chem.* 72 (2000) 3436–3442.
- [15] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, *Electrophoresis* 20 (1999) 3551–3567.
- [16] K.R. Clauser, P. Baker, A.L. Burlingame, *Anal. Chem.* 71 (1999) 2871–2882.
- [17] W. Zhang, B.T. Chait, *Anal. Chem.* 72 (2000) 2482–2489.
- [18] E. Krause, H. Wenschuh, P.R. Jungblut, *Anal. Chem.* 71 (1999) 4160–4165.
- [19] B. Thiede, S. Lamer, J. Mattow, F. Siejak, C. Dimmler, T. Rudel, P.R. Jungblut, *Rapid Commun. Mass Spectrom.* 14 (2000) 496–502.
- [20] S. Haebel, T. Albrecht, K. Sparbier, P. Walden, R. Korner, M. Steup, *Electrophoresis* 19 (1998) 679–686.
- [21] M. Hamdan, M. Galvani, P.G. Righetti, *Mass Spectrom. Rev.* 20 (2001) 121–141.
- [22] K.C. Parker, *J. Am. Soc. Mass Spectrom.* 13 (2002) 22–39.
- [23] B. Herbert, F. Hopwood, D. Oxley, J. McCarthy, M. Laver, J. Grinyer, A. Goodall, K. Williams, A. Castagna, P.G. Righetti, *Proteomics* 3 (2003) 826–831.
- [24] P.C. Liao, J. Leykam, P.C. Andrews, D.A. Gage, J. Allison, *Anal. Biochem.* 219 (1994) 9–20.
- [25] W.J. Welch, *J. Biol. Chem.* 260 (1985) 3058–3062.
- [26] A.P. Arrigo, W.J. Welch, *J. Biol. Chem.* 262 (1987) 15359–15369.
- [27] M. Gaestel, W. Schroder, R. Benndorf, C. Lippmann, K. Buchner, F. Hucho, V.A. Erdmann, H. Bielka, *J. Biol. Chem.* 266 (1991) 14721–14724.
- [28] B.R. Graveley, *Trends Genet.* 17 (2001) 100–107.
- [29] J. Graw, *Biol. Chem.* 378 (1997) 1331–1348.
- [30] L.H. Cohen, L.W. Westerhuis, W.W. de Jong, H. Bloemendal, *Eur. J. Biochem.* 89 (1978) 259–266.
- [31] M. Galvani, M. Hamdan, B. Herbert, P.G. Righetti, *Electrophoresis* 22 (2001) 2058–2065.
- [32] A. Krah, F. Schmidt, D. Becher, M. Schmid, D. Albrecht, A. Rack, K. Buttner, P.R. Jungblut, *Mol. Cell Proteomics* 2 (2003) 1271–1283.
- [33] F. Schmidt, A. Lueking, E. Nordhoff, J. Gobom, J. Klose, H. Seitz, V. Egelhofer, H. Eickhoff, H. Lehrach, D.J. Cahill, *Electrophoresis* 23 (2002) 621–625.
- [34] J. Mattow, U.E. Schaible, F. Schmidt, K. Hagens, F. Siejak, G. Brestrich, G. Haeselbarth, E.C. Muller, P.R. Jungblut, S.H. Kaufmann, *Electrophoresis* 24 (2003) 3405–3420.
- [35] F. Schmidt, M. Schmid, P.R. Jungblut, J. Mattow, A. Facius, K.P. Pleissner, *J. Am. Soc. Mass Spectrom.* 14 (2003) 943–956.
- [36] U.K. Laemmli, *Nature* 227 (1970) 680–685.
- [37] S. Lamer, P.R. Jungblut, *J. Chromatogr. B* 752 (2001) 311–322.