Research Article

MALDI-TOF/TOF de novo sequence analysis of 2-D PAGE-separated proteins from *Halorhodospira halophila*, a bacterium with unsequenced genome

Because protein identifications rely on matches with sequence databases, high-throughput proteomics is currently largely restricted to those species for which comprehensive sequence databases are available. The identification of proteins derived from organisms with unsequenced genomes mainly depends on homology searching. Here, we report the use of a simplified, gel-based, chemical derivatization strategy for de novo sequence analysis using a MALDI-TOF/TOF mass spectrometer. This approach allows the determination of de novo peptide sequences of up to 20 amino acid residues in length. The protocol was applied on a proteomic study of 2-D PAGE-separated proteins from *Halorhodospira halophila*, an extremophilic eubacterium with yet unsequenced genome. Using three different homology-based search algorithms, we were able to identify more than 30 proteins from this organism using subpicomole quantities of protein.

Keywords: Chemically assisted fragmentation / Guanidination / Homology search / Sulfonation DOI 10.1002/elps.200500959

1 Introduction

Proteomics is playing a pivotal role in the postgenome era in helping to define the functional role of genes. MS, hyphenated with a range of electrophoretic and multidimensional chromatographic separation techniques, has emerged as a key platform technology in proteomics for the rapid and high-throughput identification, characterization, and quantitation of proteins [1]. Typically, proteins are digested using trypsin and the resultant peptides are then subjected to MS analysis. The tryptic peptides provide a characteristic PMF which can be used to identify proteins. Although this approach is useful to identify proteins in simple mixtures, peptide sequence information obtained by MS/MS is required to identify individual proteins in more complex samples [2]. Sophisticated algorithms (e.g., SEQUEST and MASCOT) have been developed to aid in this process, starting from peptide MS/MS data whereby peptides are identified by correlating the uninterpreted MS/MS spectra with simulated (predicted) product ion spectra derived from peptides of the same mass contained in the databases. While the above-mentioned algorithms for protein identification from peptide MS/MS data have enjoyed considerable success, their utility is directly related to the quality of the product ion spectra and depends on the availability of database information about the proteins under investigation. For proteins not contained within sequence databases, it is necessary to determine partial or complete amino acid sequences using either manual or automated de novo peptide sequence analysis methods.

Since manual de novo sequencing is a very time-consuming process, several software tools were developed that deduce an amino acid sequence from an MS/MS
spectrum. Interpretation of MS/MS spectra relies on measuring the mass differences between adjacent fragment ion peaks of one of the major ion series, i.e., b-series (ions containing N-terminus) or y-series (ions containing C-terminus), which are common in tryptic peptides. However, most of de novo sequencing software tools inevitably suffer from inherent limitations of MS/MS spectral analysis, including incomplete b- and y-ion series (gaps), the presence of other peptide-derived peaks such as a-ions, internal fragments and neutral losses of water or ammonia [3]. The recent introduction of MALDI-TOF/TOF MS technology offers the advantage of MALDI ionization with MS/MS in a TOF instrument [4]. De novo sequencing of underivatized peptides using MALDI-TOF/TOF MS has recently been demonstrated by Yergey et al. [5]. However, compared to the MS/MS spectra of doubly charged ESI-generated ions, MS/MS spectra of singly charged MALDI ions contain more ions from other fragment ion series. Therefore, a number of derivatization methods have been proposed to improve the fragmentation of singly charged ions.

One approach to facilitate the interpretation of MS/MS spectra is to enrich a series of fragments by attaching a strongly negatively or positively charged group to the N-terminus of peptides. Keough et al. [6] demonstrated that the N-terminus can be derivatized by acylation with 2-sulfobenzoic acid cyclic anhydride or chlorosulfonylethacetate. N-Terminal sulfonic acid derivatives were subsequently proposed for peptide sequencing by ESI-MS, MALDI-TOF MS [7], and MALDI-TOF/TOF MS [8]. The introduction of a sulfo group facilitates the MS/MS fragmentation of singly charged peptide ions by providing a second “mobile” proton which lowers amide bond strength and allows more facile unimolecular decay [9]. Since sulfonation reagents react with amino groups, this derivatization results in the modification of both the N-termini and the ε-amines of lysine-containing peptides. Therefore, Keough et al. [10] expanded their approach and combined guanidination of lysine residues with the addition of a sulfonic acid group to the N-terminus. Following guanidination of lysine ε-amines, introduction of sulfonic acid groups to tryptic peptides is possible solely at the N-terminus. We further improved this method by performing the Lys side-chains’ modification directly on gel-separated proteins prior to tryptic digestion. In this way, removal of the molar excess of guanidination reagents can simply be accomplished during the destaining step of the gel spots [11].

Today, most of the proteomic studies of extremophilic bacteria have been performed on members of the Archaea for which a sequenced genome is available. Zhu et al. [12] applied the MudPIT approach to analyze the proteome of Methanococcus jannaschii, an autotrophic methanoarchaeon and the first member of the Archaea with a completely sequenced genome [12]. A shotgun approach was also used to identify the Sulfolobus solfataricus P2 proteome, a thermo-acidophilic crenarchaeon [13]. The cytosolic and membrane proteome of Halobacterium salinarum has been analyzed using PMF and LC-MS/MS techniques [14, 15]. H. salinarum is a member of the halophilic Archaea and an important model organism to study adaptations necessary for living in salty habitats. The genome sequence of Halobacterium species NRC-1 has completely been determined [16].

The extremely halophilic purple phototrophic bacterium Halorhodospira (formerly Ectothiorhodospira) halophila shows a photophobic response toward intense blue light. The wavelength dependence of this response corresponds with the absorption spectrum of the photoactive yellow protein (PYP), which suggests this protein to be the primary photoreceptor for this response [17]. Photoreceptors allow living organisms to make optimal use of the light conditions for growth and development and/or the protection from light damage. Various types of light-induced sensory responses have been characterized physiologically in detail. However, the molecular basis of this type of response is only slowly emerging. While the PYP protein is extremely well studied at the physical level, direct proof of a link between PYP and negative phototaxis is lacking. Moreover, in other species that produce PYP, a link with phototaxis has never been reported. A major limitation to the study of the physiological function of PYP is the fact that sequence information about the genome is not available (a DOE funded sequence program is currently running) and, more in particular, genetic techniques are poorly developed in this organism. There is limited information about the flanking regions of the PYP gene [18] in H. halophila and other species, but except for the presence of the biosynthetic genes for the production of the cofactor (p-coumaric acid) and for its covalent attachment to the protein, there are no generalizations that provide clues concerning the role of PYP.

Here, as a proof of principle, we applied our improved MS identification approach to identify a number of 2-D PAGE-separated proteins from H. halophila. (Partial) Sequences of tryptic peptides were submitted to homology search for identification of the corresponding protein. For this purpose, we applied three different homology-based search algorithms: FASTS, MS-BLAST, and MS-Homology [19–21].
2 Materials and methods

2.1 Materials

Urea, ammonium persulfate, CBB G-250, and agarose were obtained from Amersham Biosciences (Uppsala, Sweden). Iodoacetamide, CHAPS, DTT, and TEMED were from Fluka (Buchs, Switzerland). IPG strips, SDS, glycerine, and ampholytes were purchased from BioRad (Hercules, CA, USA). The acrylamide/bisacrylamide solution was obtained from National Diagnostics (Atlanta, GE, USA), and the solvents for mass spectrometric sample preparation were from Biosolve (Valkenswaard, The Netherlands).

2.2 Bacterial growth and preparation of extracts

H. halophila SL-1 was grown anaerobically under tungsten illumination at 30°C in medium 253 described by DSMZ, and 2.5 mL of these cultures were used to inoculate 250 mL anaerobically prepared medium. Cultures were grown anaerobically under tungsten illumination or green and blue light conditions at 30°C and harvested at the late-exponential growth phase. After washing with dH2O, the cells were resuspended in 100 mM Tris-HCl, pH 8.0, supplemented with 50 μg DNAse and 0.5 mM PMSF, and fractionated by sonication, followed by centrifugation to remove the cell debris (14 000 rpm, 30 min, 4°C).

2.3 2-DE and analysis

After determination of the protein concentration with the Protein Assay Kit (BioRad), approximately 250 μg of protein was mixed with IPG rehydration buffer (8 M urea, 2% w/v CHAPS, 0.3% DTT, final volume = 360 μL). The strips were allowed to rehydrate for 7 h and to focus (IEF) using a Multiphor II system (Amersham Biosciences) running the following program: 150 V (30'), 150 V (120'), 300 V (30'), 300 V (45'), 3500 V (90'), 3500 V (540'), 500 V (10'), and held at 500 V. The temperature was kept at 18°C. After completion of the IEF program, the IPG strips were equilibrated in a 50 mM Tris-HCl solution, pH 8.8, containing 6 M urea, 30% glycerol, 2% SDS, and 1% DTT, for 10 min, after which the solution was replaced with the same solution, except that DTT was exchanged by 5% iodoacetamide. The strips were then placed on the home-casted vertical SDS-PAGE gels and subjected to electrophoresis at 10 mA/gel for 15 min, followed by a ±5 h run at 20 mA/gel until the bromophenol blue front reached the bottom of the gel. Staining was performed using CBB G-250. The 2-D-gel images were digitized using a GS-710 densitometer (BioRad) and analyzed with the accompanying PDQuest 7.1 software (BioRad).

2.4 In-gel guanidination

Guanidination was performed by adding 5 μL of Milli-Q water, 11 μL of 7 N ammonium hydroxide (Merck, Darmstadt, Germany), and 3 μL of a freshly prepared 7.5 M O-methylisourea hemisulfate solution (Acros, Geel, Belgium) to the excised spots. The samples were vortexed briefly and incubated at 65°C. After incubation for 2 h, the guanidinated samples were taken from the oven and the remainder of the solution was discarded. The gel pieces containing the guanidinated samples were desalted and destained in one step. Two washes using 150 μL of 200 mM ammonium bicarbonate in 50% ACN/MQ (30 min at 30°C) were performed and subsequently the gel pieces were dried in a SpeedVac (Thermo Savant, Holbrook, NY).

2.5 Trypsin digestion and sulfonation

A volume of 8 μL digestion buffer (50 mM ammonium bicarbonate, pH 7.8) containing 150 ng modified trypsin per microliter (Promega, Madison, WI) was added to the dried gel spots and the tubes were kept on ice for 45 min to allow the gel pieces to be completely soaked with the protease solution. Digestion was performed overnight at 37°C, the supernatants were recovered and the resulting peptides were extracted twice with 35 μL of 60% ACN/0.1% DIEA. The extracts were pooled and dried in the SpeedVac. The peptides were redissolved in 4 μL of 12.5 mM ammonium bicarbonate and 50% ACN/MQ, and 2 μL was mixed with 2 μL of the sulfonation solution. The sulfonation reagent was prepared by dissolving 2 mg 2-sulfobenzoic acid cyclic anhydride in 1 mL dry THF to attain a 0.01 mM solution. The tubes were briefly vortexed and reacted for 15 min at room temperature. Upon sulfonation, the samples were not desalted, a fraction of the samples was mixed with matrix solution and spotted on the MALDI plate.

2.6 MS and MS/MS

A 4700 Proteomics Analyzer (Applied Biosystems, Foster City, CA) with TOF/TOF optics was used for all MALDI-MS and MS/MS applications. Samples were prepared by mixing 0.7 μL of the sample with 0.7 μL matrix solution (7 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) in 50% ACN containing 0.1% TFA) and spotted on a stainless steel 192-well target plate. They were allowed to air-dry at room temperature, and were then inserted in the mass spectrometer and subjected to mass analysis. The mass spectrometer was externally calibrated with a mixture of angiotensin I, Glu-fibrino-peptide B, ACTH (1–17), and ACTH (18–39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide B.
All of the sulfonated peptides were subjected to MS/MS using a MALDI-TOF/TOF instrument. In an initial study, using this method, in which the fragmentation spectra resulting from high- and low-energy CID experiments were compared, the authors concluded that the difference in fragmentation and the effect on database search results were surprisingly small [22]. The major difference observed is the presence of high-energy fragment ions (w-ions) in the high-energy CID spectra of some peptides. When the CID mode (gas on, collision energy 0.5 to >3.5 keV) is applied, a larger number of low-molecular weight fragments (immonium ions, internal fragments) have been observed [23]. However, it has also been shown that the use of high-energy CID results in a loss of sequence information, as the y-ion abundance decreases at both higher gas pressure and higher collision energy [24]. Therefore, we performed all fragmentation experiments with the collision energy set at 1 keV and no gas in the collision chamber (low-energy CID).

2.7 Database searches

De novo determined peptide sequences were deduced manually and used for similarity searches using the FASTS, MS-BLAST, and the MS-Homology algorithm. On-line submissions were performed using MS-BLAST at the Heidelberg server (http://dove.embl-heidelberg.de/ Blast2/msblast.html). Searches were performed against the nonredundant database (nrdb) using standard settings. The FASTS algorithm (http://fasta.biocch.virginia.edu/fasta_www/cgi/) was carried out using standard settings, and searches were performed against the NCBI/BLAST nrdb with BLOSUM 50 as search matrix. MS-Homology searches (Protein Prospector 4.0.5) were performed on the UCSF server against the NCBI nrdb using BLOSUM 50 as search matrix (http://prospector.ucsf.edu/ucsfhtml4.0/mshomology.htm).

The software used for similarity searches does not discriminate between the isobaric amino acids Ile and Leu. Therefore, all mass increments of 113 Da between consecutive y-ions were arbitrarily designated as Ile. The FASTS search results were considered significant if the E-value was below 1.0e-04. The MS-BLAST search results were considered significant if the resulting scores were higher than the threshold score indicated in the software. In order for a particular protein in the database to generate a hit, MS-Homology must find homologous sequences for the minimum number of peptides required to match. The scoring method used is based on a mutation matrix such as the one used in the BLAST and FASTA programs. The final score is calculated by adding the scores for the individual peptide alignments together. If there are several possible alignments of a given peptide, then the highest scoring alignment is used in the calculation. As the searches are based on similarity, proteins identified with lower scores must have the same generic function as the first hit. Proteins were considered positively identified only if all three search algorithms yielded the same homologous protein in the first hit. It has been demonstrated that indirect evidence can add to the significance of an identification [20]. Therefore, the identifications were further validated by using information such as the cleavage specificity of trypsin and sequence information resulting from known preferential fragmentation patterns of sulfonated peptides [8].

3 Results and discussion

The total protein extracts from H. halophila grown anaerobically under yellow or green/blue light were separated by 2-D PAGE. From these two gels, 100 spots were randomly selected and manually excized. The proteins were guanidinated in-gel and desalted/destained in one single step as described previously. Subsequently, the guanidinated proteins were enzymatically cleaved with trypsin and, after extraction, the peptides were sulfonated [11]. For 74 spots, we observed a good PMF of the sulfonated peptides, suitable for de novo MALDI MS/MS analysis. For the other 26 spots, we observed none or a very weak PMF, with signal intensities that were too low for MALDI MS/MS fragmentation. Previous experiments have indicated that sulfonic acid-derivatized peptides have poorer positive-ion sensitivity than the corresponding native peptides. The introduction of a negative charge usually leads to a decrease or even loss in signal intensity in positive mode [8]. Keough et al. [6] also noticed a decreased intensity of sulfonated peptides in the positive mode, compared to the negative ion mode, and apparently, some peptides show no signal above the noise level [6]. Recently, it was also demonstrated that the presence of the strong negative charge of the sulfonic group can create problems for sample desalting on RP media (low yield for less hydrophobic peptides) [25]. However, in our approach, guanidination is performed in-gel, and therefore, an additional desalting step to remove the excess of reagents can be omitted [11].

All fragmentation experiments were performed with the collision energy set at 1 keV and no gas in the collision chamber (low-energy CID). Typically, the most intense peaks in the PMF were selected for MS/MS analysis. In most spots, three to six peptide sequences, with a length varying between 5 and 20 amino acids, were obtained de novo (Tables 1a and b). In all fragment spectra, an initial loss of the sulfonic acid derivative was observed.
(Δm = 184 Da). By simple manual calculation of the differences between the adjacent y-ion fragments, the amino acid sequence could readily be interpreted. As an example, the protein in spot 5 was identified as fructose-1,6-bisphosphate aldolase (Table 1a). Upon sulfonation, four peptides were subjected to MALDI MS/MS analysis. In all fragmentation spectra, except one, we observed a complete y-ion series that could easily be interpreted, facilitating de novo sequencing (Figs. 2a–d). MS/MS spectra of derivatized peptides having an internal homo-Arg (guanidinated lysine) also produced complete y-ion series (Fig. 3a). However, the y-ion series were often accompanied with (Yi–17)-ion series, due to the neutral loss of NH₃. Although this (Yi–17) series is seen in the fragment spectra up to the step where the basic residue is cleaved off, its presence does not interfere with the sequence interpretation (Figs. 2a–d). Please note that the loss of the sulfonic group (−184 Da) and the neutral loss of ammonia (−17 Da) is summed as an initial loss of −201 Da.

During PSD experiments with derivatized peptides enhanced fragmentation at Pro residues and a reduced abundance in fragmentation at the C-terminal side of Pro has been observed [15]. Here, and in a previous study, TOF/TOF analysis of peptides containing an internal Pro indicated that y-ions resulting from cleavage on the N-terminal side of Pro are enhanced while y-ions resulting from cleavage on the C-terminal side of Pro are less abundant or almost completely depleted [8]. Therefore, peptides with a Pro residue near the N-terminus of a peptide can limit the amount of sequence information and cause gaps in the derived sequence (Figs. 2d, 3c). If the Pro residue occurs in the middle or near the C-terminus of the peptide this effect is less detrimental (Fig. 2c), and even for derivatized peptides containing internal homo-Arg and Pro residues we were able to derive uninterrupted peptide sequences of 20 amino acids (Fig. 4b and c). As proteomic strategies are becoming increasingly reliant on the use of automated database search algorithms, incorporation of “fragmentation rules”, such as the observed preferential cleavage at Xxx-Pro peptide bonds, into the database search algorithms will aid in the development of more effective tools for high-throughput protein identification. Furthermore, the occurrence of “nonsequence” specific ion fragments, such as the neutral loss of ammonia from peptides with internal Arg or homo-Arg, can be used to improve predictive models of peptide fragmentation for de novo sequence analysis.

<table>
<thead>
<tr>
<th>Spot</th>
<th>Protein Identification</th>
<th>Pept.</th>
<th>AA</th>
<th>E-score</th>
<th>MS-Blast score</th>
<th>MS-Hom. score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>Porin, Gram-negative type</td>
<td>5</td>
<td>50</td>
<td>3.50e-05</td>
<td>135</td>
<td>124</td>
</tr>
<tr>
<td>3</td>
<td>S-Adenosylmethionine synthetase</td>
<td>6</td>
<td>54</td>
<td>2.20e-17</td>
<td>202</td>
<td>216</td>
</tr>
<tr>
<td>4</td>
<td>Elongation factor Tu</td>
<td>6</td>
<td>73</td>
<td>2.90e-29</td>
<td>336</td>
<td>376</td>
</tr>
<tr>
<td>5</td>
<td>Fructose-1,6-bisphosphate aldolase</td>
<td>4</td>
<td>32</td>
<td>6.60e-14</td>
<td>115</td>
<td>186</td>
</tr>
<tr>
<td>6</td>
<td>Phosphoribulokinase</td>
<td>4</td>
<td>33</td>
<td>5.40e-05</td>
<td>95</td>
<td>129</td>
</tr>
<tr>
<td>7</td>
<td>Sarcosine-dimethylglycine methyltransferase</td>
<td>4</td>
<td>31</td>
<td>0.00021</td>
<td>–</td>
<td>113</td>
</tr>
<tr>
<td>8</td>
<td>Acetoacetyl-CoA reductase</td>
<td>4</td>
<td>51</td>
<td>4.30e-10</td>
<td>149</td>
<td>159</td>
</tr>
<tr>
<td>9</td>
<td>Triosephosphate isomerase</td>
<td>3</td>
<td>29</td>
<td>7.80e-12</td>
<td>156</td>
<td>121</td>
</tr>
<tr>
<td>10</td>
<td>Adenylate kinase</td>
<td>3</td>
<td>31</td>
<td>1.10e-15</td>
<td>177</td>
<td>191</td>
</tr>
<tr>
<td>11</td>
<td>2,3,4,5-Tetrahydropyridine-2,6-di-carboxylate N-succinyltransferase</td>
<td>5</td>
<td>45</td>
<td>1.40e-08</td>
<td>167</td>
<td>144</td>
</tr>
<tr>
<td>12</td>
<td>Pentose-5-phosphate-3- epimerase</td>
<td>3</td>
<td>29</td>
<td>0.00046</td>
<td>105</td>
<td>98</td>
</tr>
<tr>
<td>13</td>
<td>Superoxide dismutase</td>
<td>8</td>
<td>67</td>
<td>5.70e-17</td>
<td>313</td>
<td>298</td>
</tr>
<tr>
<td>14</td>
<td>DsrC</td>
<td>1</td>
<td>13</td>
<td>1.60e-07</td>
<td>99</td>
<td>82</td>
</tr>
<tr>
<td>15</td>
<td>Nucleoside diphosphate kinase</td>
<td>3</td>
<td>49</td>
<td>4.00e-29</td>
<td>279</td>
<td>259</td>
</tr>
<tr>
<td>16</td>
<td>Ribulose biphosphate carboxylase small chain</td>
<td>3</td>
<td>34</td>
<td>1.10e-06</td>
<td>107</td>
<td>113</td>
</tr>
<tr>
<td>17</td>
<td>Pterin-4-carboxylaminde dehydratase</td>
<td>4</td>
<td>64</td>
<td>4.00e-13</td>
<td>188</td>
<td>132</td>
</tr>
<tr>
<td>18</td>
<td>Cold-shock protein, DNA-binding</td>
<td>2</td>
<td>21</td>
<td>2.70e-08</td>
<td>112</td>
<td>115</td>
</tr>
</tbody>
</table>
Table 1b. Identified proteins from *Halorhodospira halophila* (green/blue light)

<table>
<thead>
<tr>
<th>Spot</th>
<th>Protein[bc]</th>
<th>Identification FASTS[c]</th>
<th>Pept.[d]</th>
<th>AA[e]</th>
<th>E-score[f]</th>
<th>MS-Blast score[g]</th>
<th>MS-Hom. score[h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77166263</td>
<td>Chaperone protein dnaK (Hsp70)</td>
<td>4</td>
<td>35</td>
<td>1.60e-11</td>
<td>186</td>
<td>155</td>
</tr>
<tr>
<td>2</td>
<td>54294307</td>
<td>30S ribosomal protein S1</td>
<td>3</td>
<td>34</td>
<td>1.60e-13</td>
<td>152</td>
<td>181</td>
</tr>
<tr>
<td>3</td>
<td>53762519</td>
<td>Chaperonin GroEL (HSP60 family)</td>
<td>6</td>
<td>83</td>
<td>8.60e-26</td>
<td>383</td>
<td>299</td>
</tr>
<tr>
<td>4</td>
<td>71899446</td>
<td>ATP synthase F1, alpha subunit</td>
<td>2</td>
<td>35</td>
<td>1.00e-10</td>
<td>152</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>71550918</td>
<td>Ribulose-bisphosphate carboxylase</td>
<td>4</td>
<td>30</td>
<td>1.60e-05</td>
<td>99</td>
<td>160</td>
</tr>
<tr>
<td>6</td>
<td>37527547</td>
<td>S-Adenosylmethionine synthetase (methionine adenosyltransferase) (AdoMet synthetase)</td>
<td>6</td>
<td>56</td>
<td>1.00e-22</td>
<td>194</td>
<td>257</td>
</tr>
<tr>
<td>7</td>
<td>56461311</td>
<td>Fructose/tagatose bisphosphate aldolase</td>
<td>5</td>
<td>29</td>
<td>4.40e-11</td>
<td>109</td>
<td>156</td>
</tr>
<tr>
<td>8</td>
<td>33862805</td>
<td>Phosphoribulokinase</td>
<td>3</td>
<td>36</td>
<td>1.40e-16</td>
<td>158</td>
<td>191</td>
</tr>
<tr>
<td>9</td>
<td>33152351</td>
<td>NADH-dependent enoyl-ACP reductase</td>
<td>4</td>
<td>40</td>
<td>1.20e-12</td>
<td>161</td>
<td>190</td>
</tr>
<tr>
<td>10</td>
<td>2497482</td>
<td>Adenylate kinase (ATP-AMP transphosphorylase)</td>
<td>4</td>
<td>37</td>
<td>3.1e-09</td>
<td>174</td>
<td>195</td>
</tr>
<tr>
<td>11</td>
<td>52006362</td>
<td>Dissimilatory sulfite reductase</td>
<td>1</td>
<td>13</td>
<td>0.00097</td>
<td>73</td>
<td>70</td>
</tr>
<tr>
<td>12</td>
<td>77866837</td>
<td>Nucleoside diphosphate kinase</td>
<td>2</td>
<td>24</td>
<td>4.70e-10</td>
<td>101</td>
<td>132</td>
</tr>
<tr>
<td>13</td>
<td>1402737</td>
<td>Major cold shock protein</td>
<td>1</td>
<td>11</td>
<td>9.70e-05</td>
<td>77</td>
<td>71</td>
</tr>
</tbody>
</table>

a) Spot number according to the position on the 2-D PAGE (Fig. 1).
b) NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/).
c) Identification based on FASTS search result.
d) Number of peptide sequences used in the query.
e) Total number of amino acids used in the query.
f) In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI nonredundant protein database N ~ 2075116.
g) MS-BLAST score for searches against the nonredundant protein database at http://dove.embl-heidelberg.de/Blast2/msblast.html.
h) MS-Homology (Protein Prospector 4.0.5) score against the NCBI nonredundant protein database.

**Figure 1.** 2-D PAGE-separated proteins from *H. halophila* grown anaerobically under tungsten illumination (a) or green and blue light (b). Spots in which the protein was positively identified are numbered according to Table 1.
Figure 2. MALDI MS/MS spectra (positive ion mode) of the in-gel guanidinated and sulfonated tryptic peptide mixture of spot 5 (Fig. 1). Fragment spectra of peptide IKEIHQR (a), KVNIDTDIR (b), KYAGPVFPR (c), and IKPIAIE (d). All labeled fragment ions are y-ions, (y-17)-ions resulting from the neutral loss of NH₃ are indicated as #. Where appropriate, other fragment ions are indicated (b-ions). De novo derived sequence information is indicated in the one-letter code (homoarginine (hR)). Loss of the sulfonation label is indicated as 184 Da or as a loss of 201 Da including the neutral loss of ammonia (17 Da).

The de novo determined peptide sequences were used to identify the proteins by sequence similarity searching, as reviewed by Liska and Shevchenko [26]. Given the size and growth of the current databases, it is possible that many proteins already have homologs in a database.

Database searching with MS-derived de novo peptide sequences allows the proteomic identification of proteins from organisms whose genomes have not been sequenced. However, MS and sequence similarity searches are difficult to combine. Conventional database search algorithms like BLAST or FASTA are optimized for accurate sequence queries that are longer than 35 amino acid residues. Usually, peptide sequences obtained by MS/MS do not exceed the length of a tryptic peptide, typically comprising 10–15 amino acids and, therefore, the statistical significance of retrieved hits is often ambiguous. Several database searching approaches have been reported that accommodate specific requirements of MS/MS sequencing. MS-driven BLAST (MS-
Electrophoresis 2006, 27, 2702–2711
Proteomics and 2-DE

Figure 4. MALDI MS/MS spectra (positive ion mode) of the in-gel guanidinated and sulfonated tryptic peptide mixture of spot 15 (Fig. 1). Fragment spectra of peptide FFFKA-DEICSR (a), EIMGATNPKEAAGTIR (b), and TISIIKPDAVAQNAIGEIAR (c). Labeling is as in legend Fig. 2.

BLAST) is a database search protocol for identifying unknown proteins by sequence similarity to homologous proteins available in a database. MS-BLAST utilizes redundant, degenerate, and partially inaccurate peptide sequence data obtained by de novo interpretation of MS/MS spectra. MS-BLAST does not allow gaps within individual peptides, while gaps between peptides are not penalized and can be of arbitrary length. Therefore, all peptide sequences obtained by the interpretation of acquired MS/MS are assembled into a single searching string in arbitrary order [20, 27]. MS-Homology is a database searching tool from the UCSF Mass Spectrometry Facility (Protein Prospector 4.0.5) that performs homology-based searches [21]. The program allows the comparison of a number of de novo derived peptide sequences, followed by the maximum number of amino acid substitutions allowed for each sequence, against a selected database. Different peptides from the same unknown protein can be entered in the list. A database search will look for proteins containing peptides identical
or homologous to the listed sequences. The quality of the results will be dependent on the number of peptides sequenced and the accuracy of the sequence information entered, as well as on database completeness and species to species sequence variability for the peptides entered. It is also possible to enter a part of the sequence as a mass, along with a tolerance factor. FASTS is a recently reported sequence similarity search algorithm designed to use de novo sequence data from organisms lacking comprehensive sequence data. FASTS searches databases using peptide sequences of unknown order, evaluating all possible arrangements of the peptides. The algorithm uses the heuristic FASTA comparison strategy to accelerate the search, but also uses alignment probability, rather than a similarity score, as the criterion for alignment optimality. Because the true order of the query peptides used by FASTS is not known, FASTS only requires that the aligned peptides do not overlap [19].

The de novo derived sequence information from each spot was combined in one search query and analyzed using the three search algorithms. Only when the top results (first hits) from the three searches yielded the same protein, the identification was considered as positive.

The identification of proteins was often prohibitively challenging because of variation in amino acid side chains and their relative order in a peptide backbone, and the presence of side-products such as backbone, and the presence of side-products such as neutral loss ions, contaminants, or noise peaks. Improvement of the fragmentation efficiency of peptides is of particular importance for MALDI-generated ions, because the predominant singly charged ions in MALDI generally fragment less good than doubly charged ions. The approach demonstrated here, consisting of de novo sequence analysis of derivatized peptides and homology-based identification, is a powerful technique for the identification of proteins with no genomic or other database information.

4 Concluding remarks

The rapid and accurate identification of proteins is the primary goal of modern proteomics. MS/MS can generate some useful sequence information. However, manual interpretation of peptide spectra for de novo sequencing is often prohibitively challenging because of variation in favored ion fragmentation sites, the chemical nature of amino acid side chains and their relative order in a peptide backbone, and the presence of side-products such as neutral loss ions, contaminants, or noise peaks. Improvement of the fragmentation efficiency of peptides is of particular importance for MALDI-generated ions, because the predominant singly charged ions in MALDI generally fragment less good than doubly charged ions. The approach demonstrated here, consisting of de novo sequence analysis of derivatized peptides and homology-based identification, is a powerful technique for the identification of proteins with no genomic or other database information.

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

www.electrophoresis-journal.com
For 75% of the spots we observed a PMF upon in-gel guanidination and sulfonation of the extracted peptides. An apparently bad feature of the sulfonic acid derivatized peptides is its lower intensity in positive mode analysis, partly due to the suppression effect of the strong negative charge from the sulfonic group. The poorer positive ion-sensitivity is counterbalanced by a far more efficient fragmentation of sulfonated compared with non-sulfonated peptides. TOF/TOF analysis of underderivatized peptides typically results in complex fragment spectra. After guanidination and sulfonation, a contiguous series of y-ions was observed in almost all of the fragmentation spectra (Supplementary Table 1). The y-ion series could easily be interpreted (manually or by using an algorithm) facilitating de novo sequencing.

The occurrence of “nonsequence” specific ion fragments, such as the neutral loss of ammonia, and preferential fragmentation pathways, such as the Xxx-Pro bond, can be used to improve predictive models of peptide fragmentation for de novo sequence analysis. The current understanding of the fragmentation mechanisms is still insufficient to ensure a high correlation between theoretically predicted MS/MS spectra and experimental results.

MS-Homology, MS-BLAST, and FASTS methods provide independent means of evaluating the statistical significance of hits and, therefore, it is not necessary to compare retrospectively the matched peptide sequences with actual tandem mass spectra to rule out false positive hits. As previously reported, the peptide sequences in the query were arbitrarily chosen [19, 28]. However, in this study we observed that the identification scores vary, according to the position of the sequences in the query, when applying the MS-BLAST and FASTS algorithm, a contradiction that will be the subject of a further study.

B.S. is a Postdoctoral Fellow of the Fund for Scientific Research-Flanders (F.W.O.-Vlaanderen, Belgium). K.S. is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (I.W.T.-Vlaanderen). The authors would like to thank L. Vandermeersch for excellent technical assistance.

5 References