Iterative Data Analysis is the Key for Exhaustive Analysis of Peptide Mass Fingerprints from Proteins Separated by Two-Dimensional Electrophoresis

Frank Schmidt, Monika Schmid, and Peter R. Jungblut
Core Facility Protein Analysis, Max Planck Institute for Infection Biology, Berlin, Germany

Jens Mattow
Department of Immunology, Max Planck Institute for Infection Biology, Berlin, Germany

Axel Facius
Center for Environment and Health, Institute for Bioinformatics, GSF, Berlin, Germany

Klaus-Peter Pleissner
Core Facility Bioinformatics, Max Planck Institute for Infection Biology, Berlin, Germany

Peptide mass fingerprinting (PMF) is a powerful tool for identification of proteins separated by two-dimensional electrophoresis (2-DE). With the increase in sensitivity of peptide mass determination it becomes obvious that even spots looking well separated on a 2-DE gel may consist of several proteins. As a result the number of mass peaks in PMFs increased dramatically leaving many unassigned after a first database search. A number of these are caused by experiment-specific contaminants or by neighbor spots, as well as by additional proteins or post-translational modifications. To understand the complete protein composition of a spot we suggest an iterative procedure based on large numbers of PMFs, exemplified by PMFs of 480 *Helicobacter pylori* protein spots. Three key iterations were applied: (1) Elimination of contaminant mass peaks determined by MS-Screener (a software developed for this purpose) followed by reanalysis; (2) neighbor spot mass peak determination by cluster analysis, elimination from the peak list and repeated search; (3) re-evaluation of contaminant peaks. The quality of the identification was improved and spots previously unidentified were assigned to proteins. Eight additional spots were identified with this procedure, increasing the total number of identified spots to 455. (J Am Soc Mass Spectrom 2003, 14, 943–956) © 2003 American Society for Mass Spectrometry

The development of high-resolution two-dimensional electrophoresis (2-DE) [1, 2] for protein separation and highly sensitive methods for protein identification, most notably N-terminal sequencing [3, 4] and mass spectrometry (MS) [5–7], represent major milestones on the way to an understanding of the proteome. A further step was the application of mass spectrometry for peptide mass fingerprinting (PMF) instead of SDS gel electrophoresis [8], which resulted in a quick and reliable identification of 2-DE separated protein spots [9, 10]. Instead of measuring the mass of the uncleaved proteins the high mass accuracy of MALDI- and ESI-MS in the Mr range between 500 and 3000 was utilized.

PMF as a probability method is especially powerful if the data set of potential protein candidates is small, as in the case of microorganisms with completely sequenced genomes. For unequivocal identification of post-translational modifications, complementation by protein sequencing methods is necessary. PMF has been used successfully for the identification of several hundred proteins of *Haemophilus influenzae* [11], *Mycobacterium tuberculosis* [12], *Helicobacter pylori* [13], and *Mycoplasma pneumoniae* [14]. These investigations clearly demonstrated that proteins often occur as different protein species at different positions within the 2-DE gel. However, one spot may also represent several

protein species or even several proteins. We were able to identify the main components of 2-DE separated spots with success rates of about 90%. Some of the problems that need to be addressed to fully utilize the information content of a 2-DE pattern are: (1) Identification of the remaining spots, (2) the discrepancy between sensitivity of MS and protein amount in 2-DE spots, (3) the structural discrimination between protein species, and (4) the assignment of all mass peaks in PMFs.

H. pylori with a genome of about 1600 genes is an ideal model organism for proteomics. In a first attempt we were able to identify 152 protein spots from high-resolution 2-DE gels [13] with PMF and constructed a 2D-PAGE database which is available at http://www. mpiib-berlin.mpg.de/2D-PAGE. In-gel digestion, use of a volatile buffer and advances in MS technology allowed us to identify proteins with greater sensitivity and mass accuracy [15]. Now, as a rule of thumb all Coomassie Brilliant Blue G250 (CBB G250) stained spots of a microorganism with a completely sequenced genome can be identified by PMF. Sequence coverage (SC) of at least 30% was defined as a minimum for identification. Lower values were only accepted, if the three main peaks fit the most probable candidate [12, 16]. If these criteria were not fulfilled, PMF had to be complemented by sequence information. Interestingly, although it was possible to improve SC by digest optimization and mass spectrometer tuning, the number of mass peaks which could not be assigned to the identified protein increased more than the fitting mass peaks. We obtained a 100% SC for one protein (unpublished observation of our laboratory). But a complete assignment of all mass peaks of a PMF from a 2-DE gel spot became more complicated since more than 100 peaks can be detected in one given PMF (unpublished observation of our laboratory). We became aware of many chemical modifications such as methionine and tryptophan oxidation, methylation caused by CBB G250 staining, pyroglutamic acid formation, incomplete cleavages, trypsin autolysis products and contamination from keratins of human or animal origin [17]. The Mr range between 500 and 1000 additionally contains mass peaks of matrix derivatives. All of these complications depend on the experimental design. A bias for arginine-containing peptides was observed [18] and integrated into a software program (ChemApplex) for optimal scoring of PMFs together with other chemical factors, in particular the amino acid composition and the sequence of the amino acids around the cleavage site, mass accuracy and peak intensity [19].

Here, we propose an iterative procedure to eliminate as many of the mass peaks not belonging to the main components of a spot as possible. Common contaminating peaks of CBB, keratin, matrix, and trypsin were elucidated from a data set of 480 spots from a 2-DE pattern of *H. pylori* using a self-developed software tool (MS-Screener). After eliminating these contaminants, peptide mass peaks of other *H. pylori* proteins within the 2-DE pattern were searched by clustering and also

eliminated in the peptide mass list of the spots under investigation. The remaining peptide mass list facilitated identification of formerly unidentified proteins and improved identification confidence for most of the spots.

Experimental

Two-Dimensional Electrophoresis

H. pylori was grown for two days on serum plates [12] and cellular proteins were prepared for 2-DE to obtain a final concentration of 9 M urea, 70 mM DTT, 2% Servalyte 2-4, 1% CHAPS, 1 mM PMSF, 0.1 μM pepstatin, 2.1 μM leupeptin and 2.9 mM benzamidine. Proteins were separated by 2-DE combining carrier ampholyte IEF and SDS-PAGE on 23 cm $\times$ 30 cm gels [20, 21]. This technique has a resolution power of about 5000 protein species. For analytical and preparative investigations, 0.75 mm and 1.5 mm thick gels, respectively were used. For analytical gels 50–100 μg of protein and for preparative experiments up to 900 μg of protein were utilized. The samples were applied at the anodic side of the IEF gels and run under non-equilibrium conditions. The preparative gels were stained with CBB G250 or R250.

MALDI-MS Peptide Mass Fingerprinting

A data set of 480 protein spots was analyzed by MALDI-MS PMF after in-gel digestion using trypsin (porcine sequencing grade modified; Promega, Madison, WI). The in-gel digestion protocol in this data set varied, depending on the optimization status of the procedure. Some of the most intense spots were identified with a self-developed peptide collecting device [22]. Later the peptides were concentrated using ZipTip C-18 columns (Millipore, Bedford, MA). After optimization of the procedure by reduction of the trypsin concentration and introduction of the volatile buffer ammonium carbonate, the concentration procedure was omitted [15]. Alpha-cyano-4-hydroxy cinnamic acid (α-CHCA) and dehydroxy benzoic acid (DHB) were used as matrix for 250 and 230 spots, respectively. Mass spectra were recorded in the Reflectron mode of a time-of-flight MALDI mass spectrometer (Voyager Elite, Perspective Biosystems, Framingham, MA) with delayed extraction. Spectra were obtained by the summation of 256 laser shots. Under optimal conditions internal calibration of the spectra resulted in a mass accuracy of about 30 ppm. Searches with MS-FIT (UCSF Mass Spectrometry Facility, http://prospector.ucsf. edu) or Mascot (Matrix Science Ltd., UK, http://www. matrixscience.com) in the NCBInr or SwissProt database were performed with a mass tolerance of 100 ppm. The identification criteria were an SC of at least 30% and if below at least the 3 main peaks have to fit with the candidate [12, 16].

Data Acquisition and Preprocessing: MS-Screener

We have developed a Perl-script, termed MS-Screener (download area at http://www.mpiib-berlin.mpg.de/ 2D-PAGE) that facilitates the comparison of PMF peak lists. One peak list (.pkm, .dta or .txt-file extension) is used as a template that can be searched against a data set of peak lists. The search results in an unranked list of redundant masses observed between the template and each peak list of the data set. The search can be performed with a user defined mass tolerance. The idea to compare a given PMF with a large data set of PMFs was first proposed in conjunction with the concept of minimal protein identifiers [23].

In the present study we used this self-developed tool to identify and subsequently eliminate frequently occurring contaminant peak masses. For this reason we first manually generated a template comprising all peak masses in the Mr range from 500 to 4000 in 100 ppm intervals. This template was subsequently compared to the peak lists (.pkm-files) of the data set. The results were implemented into a MySQLFront database (http:// mysqlfront.venturemedia.de/). Identitical mass values were counted via My-SQL statement and transformed into an x-y plot, where the m/z values were plotted in the x-axis and the number of observed identical peaks in the y-axis. Peaks occurring in at least 5% of all spectra were defined as contaminants. In some cases, peaks observed less frequently and not fulfilling the half-decimal place rule (see below) were also regarded as contaminants, since they had previously been described in the literature or because they were detected in a matrix reference spectrum. The results of the analysis were visualized using standard software and the identified contaminants were eliminated from the peak lists using a second Perl-script.

Cluster Analysis

In order to find groups or clusters of mass spectra and to assess their similarity/dissimilarity, hierarchical agglomerative cluster analysis [24] was applied. The goal was to find relationships between the proteins within the spots. Therefore, as a preprocessing step the contaminant peaks were detected by MS-Screener and removed from peak lists using a Perl-script. The spectra were intervalized with accuracy of 100 ppm and subsequently binary encoded. The intervalization and binarization was realized using a Java-tool. Binary coding means that only the presence or absence (L/O) of peaks was taken into account, neglecting the peak intensities.

For hierarchical clustering of binary coded spectra, the binary distance metric was calculated between the binary vectors of each pair of spectra to be clustered. The hierarchical agglomerative algorithm proceeded by merging the two closest (most similar) spectra first, and then successively merging spectra or groups of spectra in order of greatest similarity. Hierarchical clustering resulted in a tree structure (dendrogram). Spectra with the highest similarity are near the distance height 0. We used the statistical programming environment R (http://www.r-project.org) with the classical multivariate R-package "mva" for the binary distance metric (R-function: dist) and hierarchical clustering (R-function: hclust) computation running on a locally installed Rweb-server.

Results and Discussion

Helicobacter pylori 26695 Proteomics

About 1800 spots from *H. pylori* 26695 cells were separated by 2-DE (http://www.mpiib-berlin.mpg.de/2D-PAGE) [13]. The data set of the present investigation consisted of 480 PMFs from 480 spots (publication in preparation). Using the 30% SC criterion 447 spots were identified. Sixteen spots had a candidate slightly missing this criterion, and 17 of the PMFs did not result in any identification. All PMFs contained numerous mass peaks not assignable to the main component of the spot. The large data set provided the possibility to determine and eliminate peaks, which occurred repeatedly in many spectra and to develop tools for a three-step analysis of PMFs originating from 2-DE spots. It has to be stressed that our data set is not parameter-free and each laboratory needs to determine frequently occurring masses independently. Here, we focus on a procedure to understand the composition of 2-DE spots. The strategy employed the elimination of contaminants and the reduction of unassigned mass values by an iterative procedure. After this data reduction the remaining mass peaks were used for a database search and secondary components were determined. This procedure is useful when no protein could be identified previously, because the number of additional mass peaks did not allow any identification. Other goals are the identification of secondary components and post-translational modifications.

Characteristics of Peptide Masses: The Half-Decimal Place Rule

It has been observed that in contrast to other molecules, for peptides the first decimal place of the mass is near the half of the first digit of mass values between 500 and 999 and near the half of the first two digits of mass values between 1000 and 1999 [25, 26]. In the mass range 2000 to 3000 again the first decimal place is near the half of the first digit of the peptide mass. In order to test this half-decimal place rule, we performed an in silico trypsin digest of all the predicted proteins of *H. pylori* 26695 and plotted the decimal places against the masses of all peptides (Figure 1a). A maximum deviation of 25% was observed in the Mr range 500 to 1000. The deviation increases to 45% in the 2500 to 3000 Mr range allowing only the exclusion of decimal places between 6 and 8. Cysteines are the main cause for deviations from the regression curve. With a mass
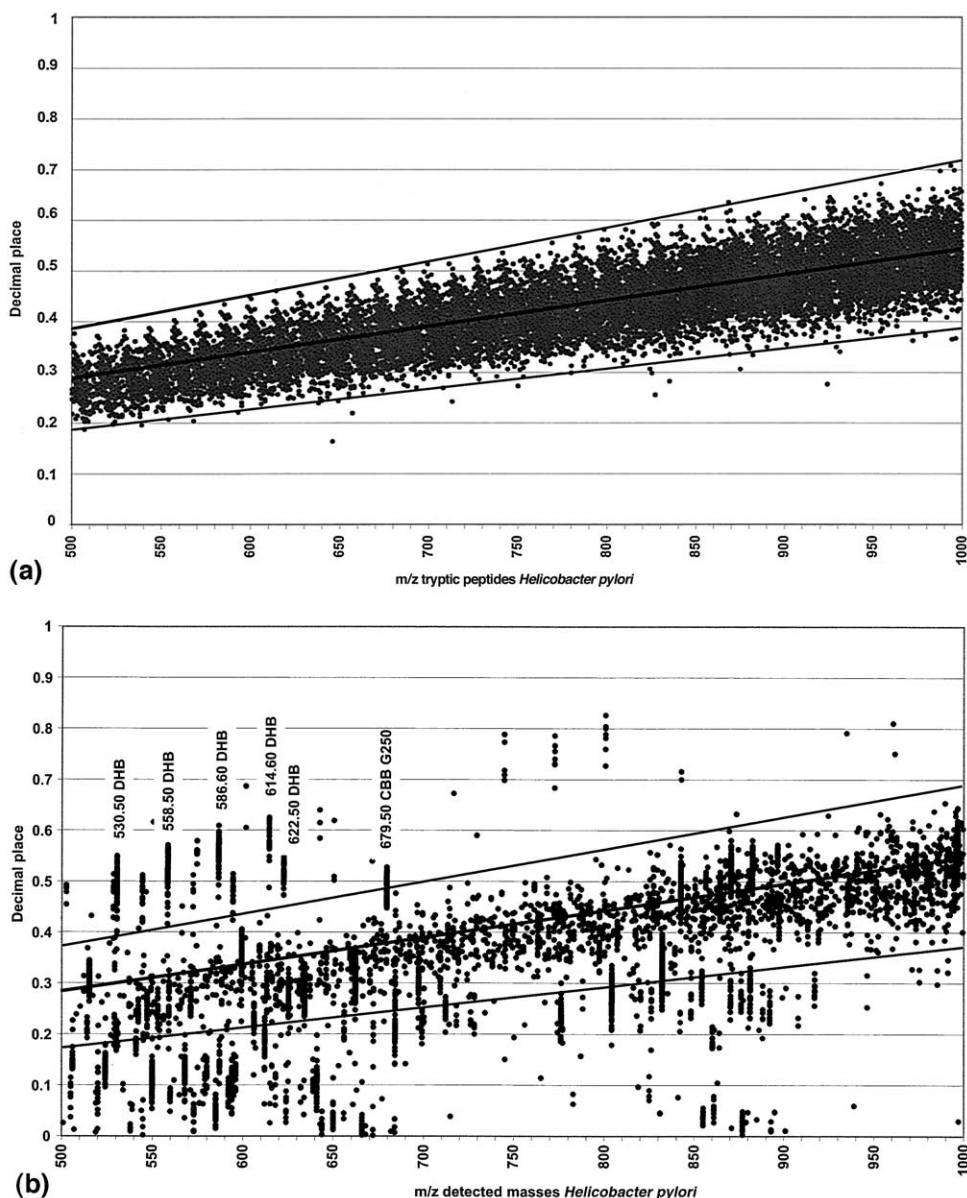
Figure 1. Half decimal place rule: Dependency of tryptic peptides Mr of *H. pylori* to the decimal of its Mr in the range from 500 to 1000 Da. (**a**) Theoretical values calculated from the sequence database of *H. pylori* proteins. The theoretical masses were generated via MS-Digest (http://prospector.ucsf.edu) and include one missed cleavage and masses with potential oxidation of methionine. With a linear regression the gradient of all values between 500 and 1000 Da was calculated. More than 99.99 % of the values were within a +/− range around the mean value at 500 and 1000 Da of 10 and 16%, respectively. (**b**) Practical values obtained from 480 PMFs from spots of a 2-DE gel. Masses of trypsin autolysis fragments, CBB G250 and matrix peaks were aggregated in small clusters, representing the mass accuracy obtained by the calibration. Masses from matrix were very often outside of the theoretical range from the decimal point.

accuracy of about 0.05 Da in this mass range it can be clearly decided if a mass value belongs to an unmodified peptide or not. Indeed, many of the observed mass values were not within this 25% tolerance range (Figure 1b), excluding them from peptide masses. In this figure, vertical lines indicate frequently detected masses with their mass deviations. The trypsin autolysis product with the mass of 842.5 appeared as a vertical line within

the expectation range for peptides, whereas the DHB derivative with the mass 530.5 was clearly outside this range. The CBB G250 mass of 832.3 illustrated that a non-peptide mass may also occur within the confidentiality range of peptides, and only peptides without modifications may be excluded for values outside this range. For large data sets this diagram may also be used for the detection of frequently observed mass values.
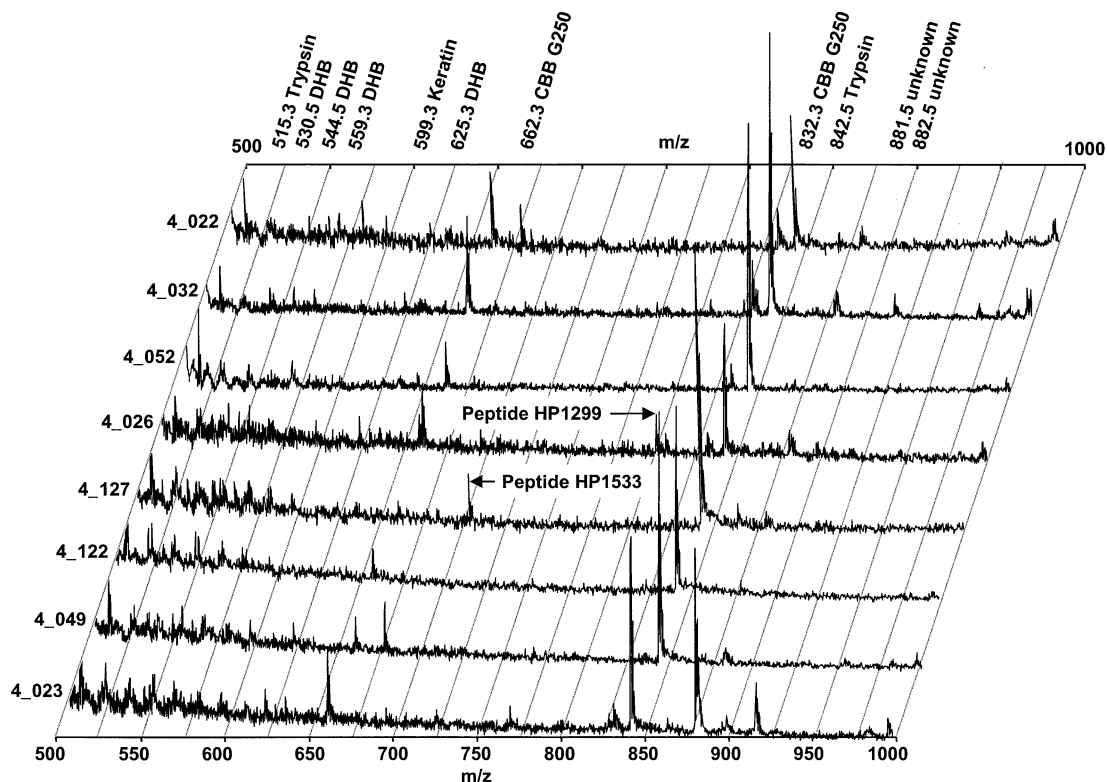
Figure 2. Example of eight MALDI mass spectra from *H. pylori* 2-DE spots in the mass range from 500 to 1000 using DHB as the matrix. Common contaminants such as trypsin autolysis products, human keratin and masses of DHB matrix cluster and two significant peptides of proteins from *H. pylori* (HP1533, 705.39 VGNIFR and HP1299, 801.5 AVILTER) are labeled.

MS-Screener: Removal of Contaminants

Figure 1b illustrates that repeatedly occurring masses may be visualized. Within this figure the number of repeats can be determined only if there is a slight shift in the measured mass. Also the stacked visualization of original mass spectra (Figure 2) did not allow the systematic detection of repeated mass values. To avoid this problem we modified the presentation of data (Figure 3). We arranged all of the masses of a PMF on a calibrated x-axis and sorted all of the PMFs one above the other. The mass values were intervalized according to their mass accuracy in 100 ppm increments. Twenty of 480 PMFs were displayed (Figure 3). These PMFs comprised a number of repeatedly occurring masses indicated as vertical lines, visualizing the MS-Screener results. Most of these masses could be designated as contaminants (Figure 4). Here, the complete data set was divided into two parts to distinguish between the contaminants of the two matrices, DHB and α-CHCA. More than 50% of the 230 DHB spectra were contaminated by the peaks 515.3 (trypsin), 530.5 (DHB), 662.3 (CBB G250), 832.3 (CBB G250), and 842.5 (trypsin) (Figure 4a). More than 50% of the 250 α-CHCA spectra were contaminated by the peaks 568.1 (α-CHCA) and 842.5 (trypsin) (Figure 4b). Obviously non-matrix contaminants are influenced by the choice of the matrix. Masses occurring in at least 5% of all PMFs were regarded as contaminants. In some cases

peaks observed in less than 5% of the PMFs were also regarded as contaminants, because they had previously been described (α-CHCA: 587.1, 641.1, 666.0, 699.2) [27] or were detected in a matrix reference spectrum. In total we elucidated 123 contaminant masses, which are listed in Table 1.

Spot Detection by Mass Spectrometry

Mass spectrometry is used in the molecular scanner [28] for detection and quantification of proteins in 2-DE gels. We applied MS detection to obtain information about the distribution of proteins within the gel. A small sector of the *H. pylori* cell protein pattern in the mass and pI-range around 53 kDa and 8.0, respectively, was analyzed in detail. Thirty pixels with a size of 3 mm × 3 mm were excised from a CBB G250 stained gel. After tryptic digestion and MALDI-MS the peak lists for each pixel were analyzed. In total, 12 proteins were recognized in this 1.5 cm × 1.8 cm large gel sector after elimination of the contaminants. The local distribution of eight of these proteins (Figure 5) clearly showed that the proteins are not restricted to a spot, rather they smear over several pixels. The clearly separated spots in the CBB G250 stained 2-DE pattern therefore represents only the tip of the iceberg. Serine protease best confirmed the spot representation with a high peak in pixel
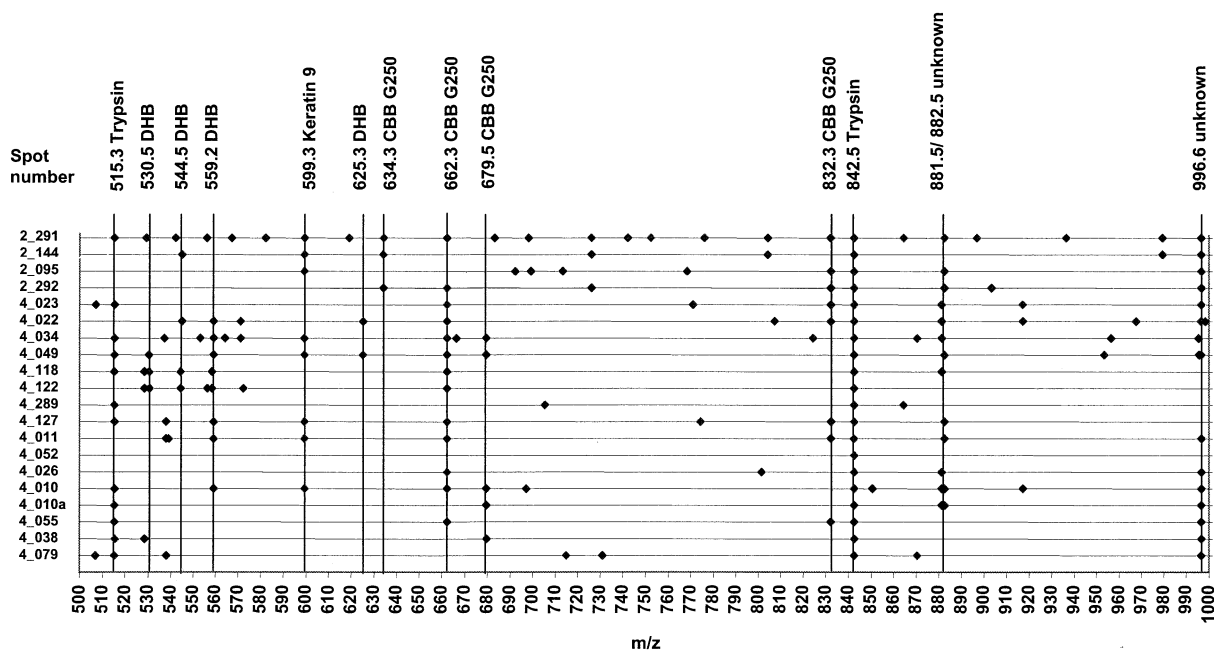
Figure 3. MS-Screener diagram of 20 PMFs using DHB as the matrix. Each mass of a PMF is represented by a dot. Contaminants occur in several spots and are labeled with vertical lines. Only masses occurring in more than 5% of all 250 PMFs were accepted as contaminant and labeled as vertical line.

10 and two lower peaks in pixel 2 and 4. Additionally, pixel 2 contains OMP20, pixel 4 OMP20 and catalase. Inosine-5′-monophosphate dehydrogenase and citrate synthase were broadly distributed over several pixels, which could not be expected from the 2-DE pattern. The long smear of malate oxidoreductase over 4 pixels was common to both detection methods. It is obvious from these results that spot visualization by staining is a simplification of the actual protein distribution within a gel.

All pixels contained peptides from more than one protein (Table 2). The total number of mass peaks ranged between 52 and 80. The 2-DE pattern predicted only one protein for pixels 10 and 16. Within these pixels distinct protein spots were observed. Nevertheless, the intensely stained pixel 10, which represents a unique spot identified as serine protease, additionally contained five peptides of 30S ribosomal protein S1 (HP0399). The results of this analysis are summarized in Table 2, which displays the pixel main components, contamination by neighbor spots, and the number of assigned peaks. Further, it includes the number of contaminant and unassigned peaks, giving an impression of the degree of contamination.

In order to analyze the relationship between the pixels with regard to their protein composition we performed a cluster analysis (Figure 6) for the gel sector shown in Figure 5. The pixels grouped in six main clusters. For this part of the gel it is obvious that the clusters were caused by horizontal protein smearing.

These observations clearly showed that after removal of contaminants in a second iteration step contamination by neighbor spots has to be considered to improve identification. Overlapping proteins in a grid of a sector of 18.3 mm × 3.7 mm were also detected in a recent study [29], where the pixel size of 0.25 mm × 0.25 mm was applied. The reproducibility of this smearing effect should be analyzed. It can be expected that the effect depends on gel thickness and amount of sample applied. Accordingly, if subtractive analysis is performed in thin silver stained gels with low amounts of sample, and identification in thick CBB stained gels with high amount of sample, the effect of neighbor spot contamination in analytical gels probably will be overestimated. Therefore, subtractive analysis using gels thicker than 1.0 mm containing large amounts of protein cannot be recommended. In contrast, high-sensitivity staining methods applicable for protein amounts of 20 to 60 μg of protein for an estimated complexity of several thousand protein species are required. Because silver staining limits the identification sensitivity of MALDI-MS [30], fluorescence detection by Sypro Ruby [31], ruthenium II tris (Bathophenanthroline disulfonate) [32], and succinimidyl esters of the cyanine dyes Cy3 and Cy5 [33] as used in difference gel electrophoresis (DIGE) [34] could present a feasible solution.

Cluster Analysis: All Spots are Connected via Peptide Masses

MS-Screener helps to detect mass values which are common to many spectra. If the relationship between different spots has to be determined cluster analysis is the method of choice. The dendrogram of this study comprised 480 mass spectra consisting of 4948 binary
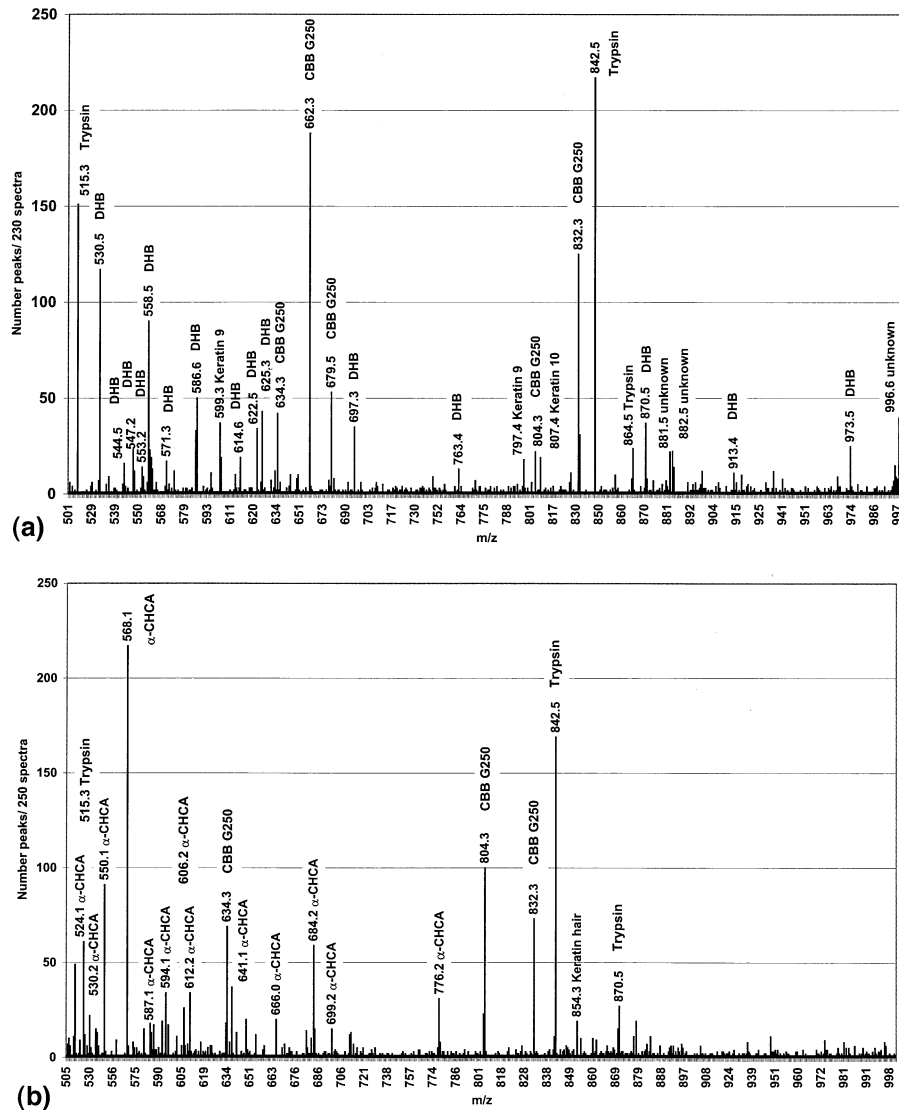
Figure 4. MALDI-MS masses from 480 PMFs of 2-DE separated spots from *H. pylori* 26695 in the mass range 500 to 1000 Da. All proteins from spots were digested with methylated porcine trypsin. With the program MS-Screener, the MALDI-MS spectra were searched for identical masses with a tolerance of 100 ppm and counted via SQL statement. Known masses of trypsin autolysis fragments, CBB G250 and most matrix peaks occurring in more than 5% of all spectra are labeled. There were no mass peaks Mr > 1000 Da for CBB G250 and matrix. (a) Data with DHB as the matrix from 230 measured spots. To visualize proteins the 2-DE gel was stained with CBB G250. (b) Data with α-CHCA as the matrix from 250 spots. Proteins were stained with CBB G250 or CBB R250.

elements. After removing masses of contaminants, a cluster analysis of all PMFs resulted in clusters between 2 and 20 spots. Remarkably, within about 480 of 1800 separated spots of a 2-DE gel from an organism with about 1500 genes there was no unique spot, indeed each spot had at least three peptides in common with another.

The question arises whether the clusters we obtained comprise spots in a narrow or widely distributed area of the gel. The cluster containing GroEL peptides exemplifies a cluster with a main spot series containing spots 1_390, 1_194, and 1_192 and a distribution of some of the mass peaks over long distances (Figure 7). Spot 2_438 reached a SC of 32% for GroEL. Therefore, the main component of this spot is GroEL. In contrast, within

spot 2_318 there are only three mass peaks of GroEL (SC 6%) and the main component here is Cag8 with a SC of 35%. In this case it seems obvious that GroEL peptides are contaminants, because the three peptide masses belong to the most intense ones of the parent spot of GroEL (1_390). The reason why this spot is contaminated and others more nearby the main GroEL spots are not, remains unclear. The MS-Screener result in the lower part of Figure 7 shows mass values which the nine spots of the cluster have in common. With the 30% SC criterion five spots (1_097, 1_192, 1_194, 1_390, and 2_438) contain GroEL as a main component. Spot 1_097 was detected at a lower Mr than the main spot of GroEL which predicts a processed form of this protein.

Table 1. Contaminants present in more than 5% of 480 PMFs. All keratins observed were of human origin.

m/z	Source	m/z	Source	m/z	Source
506.10	α-CHCA	530.50	DHB	515.30	trypsin
524.10	α-CHCA	544.50	DHB	827.50	trypsin
530.20	α-CHCA	547.20	DHB	842.50	trypsin
541.20	α-CHCA	553.20	DHB	856.50	trypsin
550.10	α-CHCA	558.50	DHB	864.50	trypsin
568.10	α-CHCA	559.20	DHB	870.50	trypsin
587.10	α-CHCA	571.30	DHB	995.50	trypsin
594.10	α-CHCA	573.00	DHB	1011.60	trypsin
606.20	α-CHCA	580.10	DHB	1033.50	trypsin
612.20	α-CHCA	585.10	DHB	1045.60	trypsin
641.10	α-CHCA	586.60	DHR	1338.80	trypsin
666.00	α-CHCA	596.10	DHB	1666.00	trypsin
684.20	α-CHCA	614.60	DHM	1940.90	trypsin
699.20	α-CHCA	622.50	DHB	1987.10	trypsin
776.20	α-CHCA	625.30	DHB	2083.00	trypsin
634.30	CBB G250	641.10	DHB	2211.10	trypsin
662.30	CBB G250	697.30	DHB	2225.10	trypsin
679.50	CBB G250	763.40	DHR	2230.20	trypsin
804.30	CBR G250	870.50	DHB	2230.30	trypsin
832.30	CBB G250	913.40	DHB	2239.10	trypsin
		973.50	DHB	2283.20	trypsin
				2297.20	trypsin
881.50	unknown			2720.30	trypsin
882.50	unknown			2807.30	trypsin
996.60	unknown			3337.70	trypsin
515.33	keratin I	547.30	keratin hair	571.27	keratin 10
547.26	keratin I	854.38	keratin hair	807.40	keratin 10
547.30	keratin I	973.50	keratin hair	995.52	keratin 10
973.53	keratin I	1107.58	keratin hair	1109.49	keratin 10
1033.52	keratin I	1179.59	keratin hair	1118.51	keratin 10
1065.53	keratin I	1302.71	keratin hair	1165.59	keratin 10
1179.60	keratin I	1323.67	keratin hair	1234.68	keratin 10
1265.64	keratin I			1300.60	keratin 10
1277.63	keratin I	599.28	keratin 9	1357.72	keratin 10
1277.71	keratin I	697.33	keratin 9	1365.64	keratin 10
1302.70	keratin I	797.35	keratin 9	1434.77	keratin 10
1302.72	keratin I	1065.50	keratin 9	1493.74	keratin 10
1357.70	keratin I	1095.55	keratin 9	1707.77	keratin 10
1383.70	keratin I	1277.59	keratin 9	2367.26	keratin 10
1393.73	keratin I	1307.68	keratin 9		
1475.75	keratin I	1323.67	keratin 9		
1475.79	keratin I	1791.73	keratin 9		
1638.86	keratin I	2221.07	keratin 9		
1716.85	keratin I	2510.13	keratin 9		
1993.98	keratin I	2705.16	keratin 9		
2383.95	keratin I				

After identifying the first spots from 2-DE gels it became obvious that one protein is not necessarily represented by one spot alone. Proteins are modified after their synthesis into different protein species and these are separated by 2-DE, if charge or Mr are changed. These protein species are distributed over the 2-DE gel and are not limited to series in a narrow Mr or pI range. This has also been observed for heat-shock proteins Hsp27 [35] in human heart and HspX [36] of *Mycobacterium tuberculosis* and as reported here for another heat-shock protein: GroEL of *H. pylori*. In addition it becomes clear that often one spot does not represent solely one protein species and even different proteins are found within one spot. All of this data

analysis so far raises many questions including: What is the relevance of common peptide masses between distant spots? Is it only a question of contamination, or are there interactions between proteins occurring during electrophoresis, or even in the native situation? May protein species with a different conformation migrate to a different position within the gel? The tools presented here: MS-Screener and cluster analysis, will now help to find answers to these questions. Sequence analysis by MS will be the method to find evidence for such hypotheses derived from PMFs.

In addition to the removal of contaminants, the removal of peptide masses of neighbor spots of the 2-DE pattern helps to increase the sensitivity for identification.
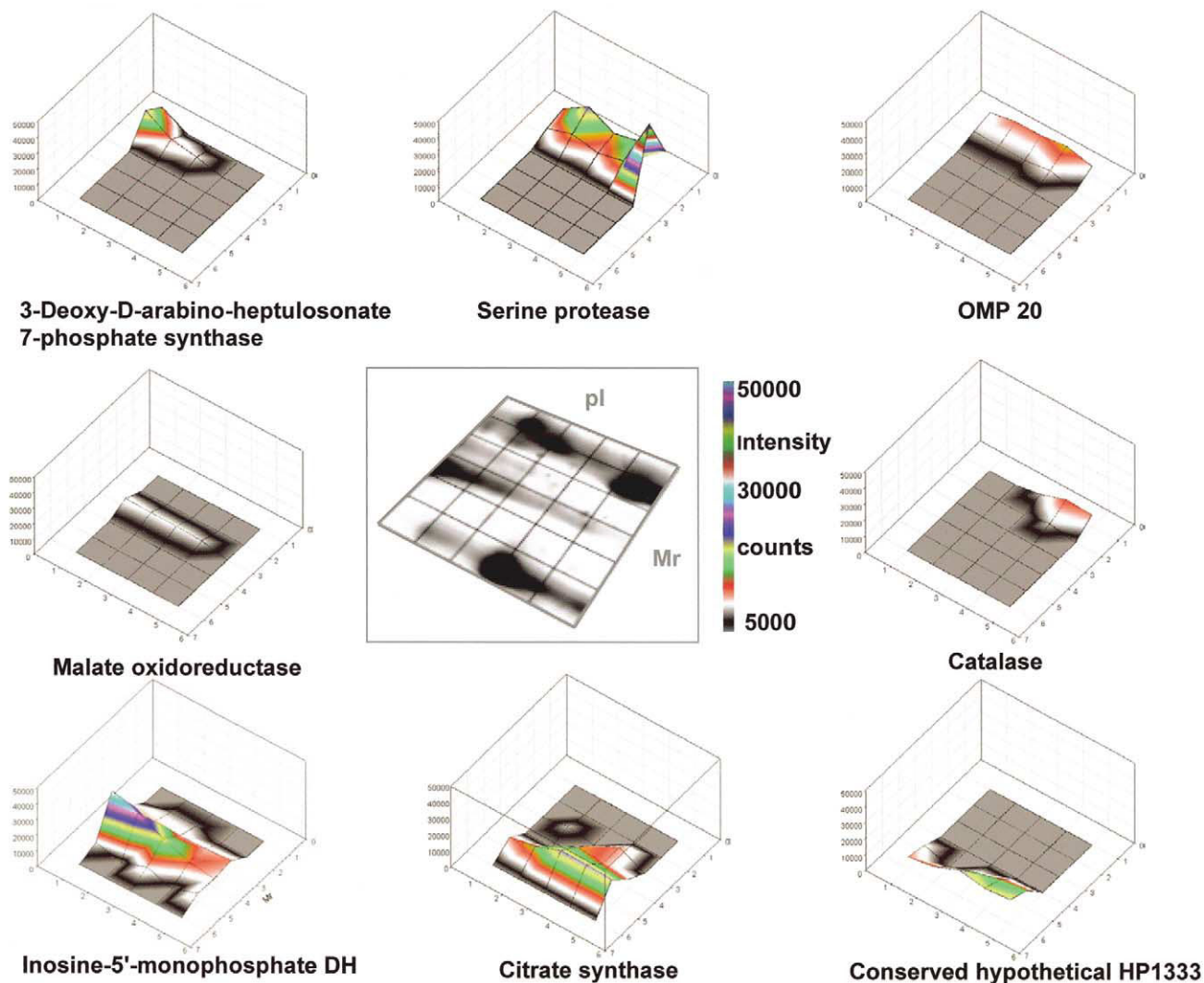
Figure 5. Distribution of proteins within a sector of the *H. pylori* 2-DE gel. Three-dimensional (3-D) mesh plots of proteins identified within the 30 pixels of the sector. The 2-DE gels were 1.5 mm thick and 440 μg *H. pylori* cell proteins were applied to the IEF gel. The sector stained with CBB G250 is shown in the center of the figure. Each pixel is analyzed by MALDI-MS PMF and only identified peptides from a protein are used in the 3-D mesh plot. For the 3-D mesh plot, the average intensity for each protein of a pixel was calculated. The average intensity was defined as the mean intensity of the three main peaks and is represented in the z-axis of each plot. The three main peaks were chosen from the spectra with the highest intensity. The x- and y-axes describe the position of the pixel. Each edge of the squares in the 3-D mesh plots represents the average intensity value of one pixel in the central stain picture and therefore in the 3-D mesh plots only twenty squares result. The 3-D mesh plots show the distribution of three main proteins, serine protease, inosine-5'-monophosphate dehydrogenase, and citrate synthase, and five low amount proteins in this sector, 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase, OMP 20, malate oxidoreductase, catalase, and the conserved hypothetical protein HP 1333.

After removal of the masses of the main components the remaining mass list may reveal further constituents of a spot, which have to be validated by MS/MS.

Procedure for Exhaustive Analysis of Peptide Mass Fingerprints

Database search algorithms are indispensable for the identification of proteins from PMFs. These tools allow the application of different search strategies depending on the PMF under investigation. In some cases, especially for faint or multi-component spots of 2-DE gels, a first search will not result in a convincing identification. In this report, we suggest an iterative procedure, which includes the knowledge of contaminants and neighbor spots comprising the complete data set (Figure 8). After elimination of contaminant and neighbor spot masses, the remaining mass list is searched again and the main component confirmed with an improved scoring value. Additional components may be identified by a subse-

Table 2. Six pixels of the 2-DE gel sector shown in Figure 4: Main components, their contaminants, neighbor spot contaminants, and the number of assigned peaks

				Number of peaks			
Pixel	Total	Contaminants	Main component	Number	Neighbor spot components	Number	Remaining
2	52	17	Serine protease HP1019	23	OMP 20 HP0912	6	3
					30S ribosomal protein S1 HP0399	2	
					Catalase HP0875	1	
					Serine protease HP1019	3	
6	57	28	3-deoxy-D-arabino-heptulosonate 7-phosphate synthase HP0134	19	OMP 20 HP0912	1	5
					Malate:quinone oxidoreductase HP0086	1	
					3-deoxy-D-arabino-heptulosonate 7-phosphate synthase HP0134	7	
7	64	25	Serine protease HP1019	16	OMP 20 HP0912	4	11
					30S ribosomal protein 20 HP0399	1	
10	60	24	Serine protease HP1019	25	30S ribosomal protein S1 HP0399	4	7
11	59	20	Malate:quinone oxidoreductase HP0086	24	Inosine-5′-monophosphate dehydrogenase HP0829	5	10
16	80	20	Inosine-5′-monophosphate dehydrogenase HP0829	36	Malate:quinone oxidoreductase HP0086	2	22
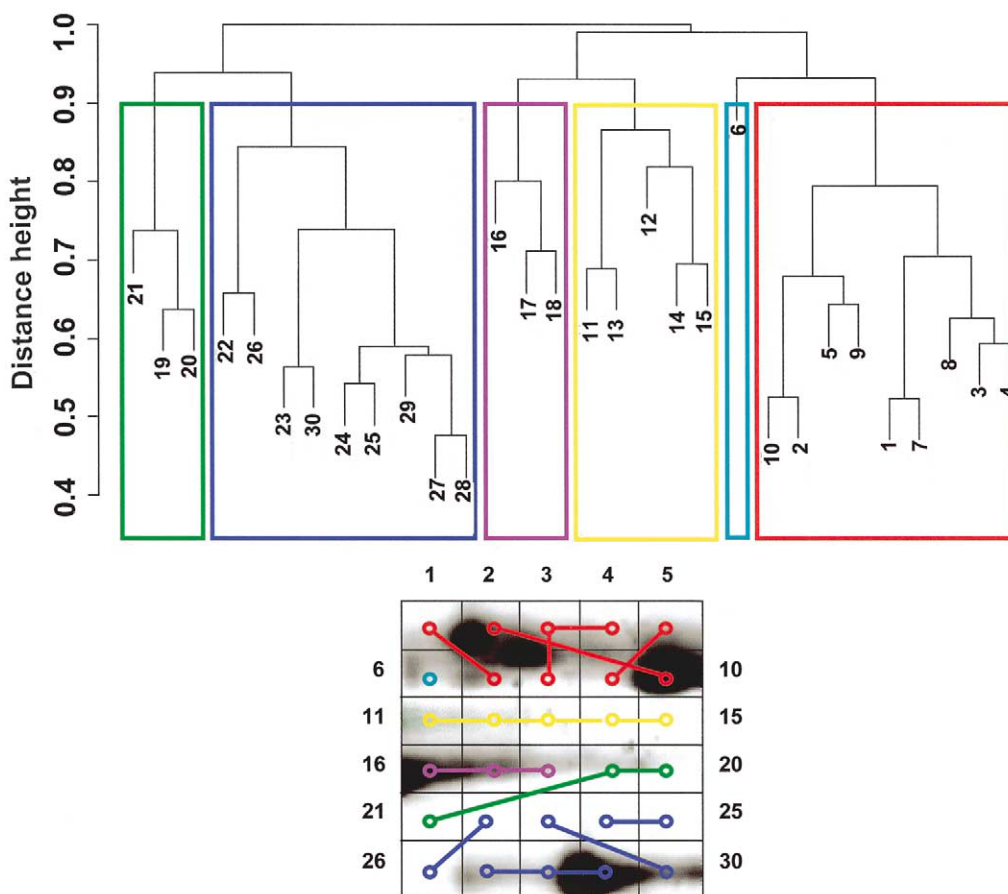
Figure 6. Cluster dendrogram of MALDI-MS spectra from 30 pixels of the 2-DE gel sector. The height of the cluster demonstrates the linkage between the spectra. Distance height between each spectrum in the data set was calculated by a binary distance metric. Six main clusters are marked with a frame. The bottom of the figure shows the 2-DE gel sector from *H. pylori* with the positions of the clusters within this sector. The sector was divided into 30 unique pixels and each pixel was analyzed by MALDI-MS PMF. The cross-links on the gel show the relationship between the spectra after clustering.
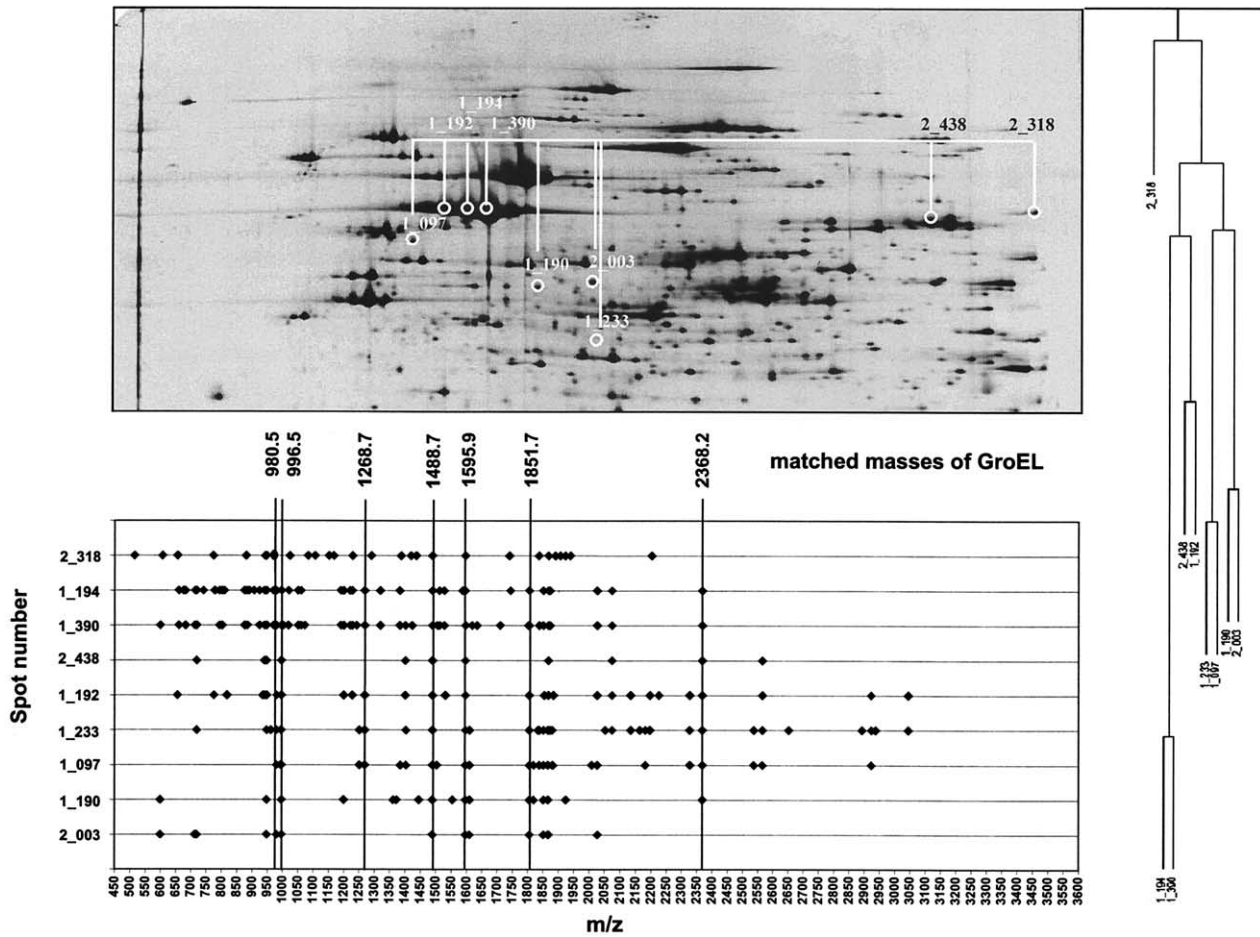
Figure 7. GroEL protein cluster within the *H. pylori* 2-DE gel. The dendrogram of the GroEL cluster is shown on the right side, the upper part of the *H. pylori* 2-DE gel with the position of the members of the GroEL cluster at the top and a MS-Screener diagram at the bottom. Nine of 480 spectra are comprising this cluster. The spectrum of spot 2_318 has the lowest similarity and only three peptides are in common with GroEL. Within the MS-Screener diagram peaks of the spectra are displayed as a quadrangle. Some of the common peaks of GroEL peptides with high intensity were connected with vertical lines, demonstrating masses responsible for cluster formation.
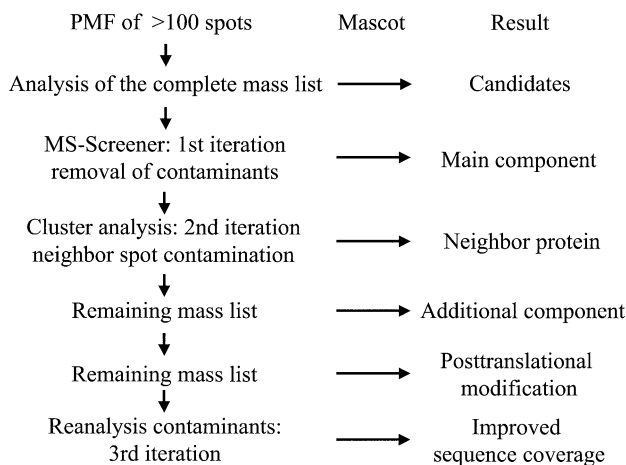

Figure 8. Flow diagram of the iterative procedure for evaluation of PMFs from 2-DE protein spots.

quent removal of masses assigned to the main component. If the masses of these additional components are removed, the chance to identify post-translational modifications increases and the remaining mass list should be searched for them. Known modifications such as methionine and tryptophan oxidation, acetylation at the N-terminus, methylation, pyro-glutamylation and others may be checked, followed by searches with the findmod tool (http://us.expasy.org/tools/findmod/). This procedure complements the prediction of post-translational modifications by iterative calibration of 2-DE gels [37]. In a third iteration the contaminant peaks should be reanalyzed to find out, whether they contain peptides of the main, neighbor spot, or additional components for further improvement of scoring factors and SC. For these considerations the peak intensity has to be included and if final doubts occur, they
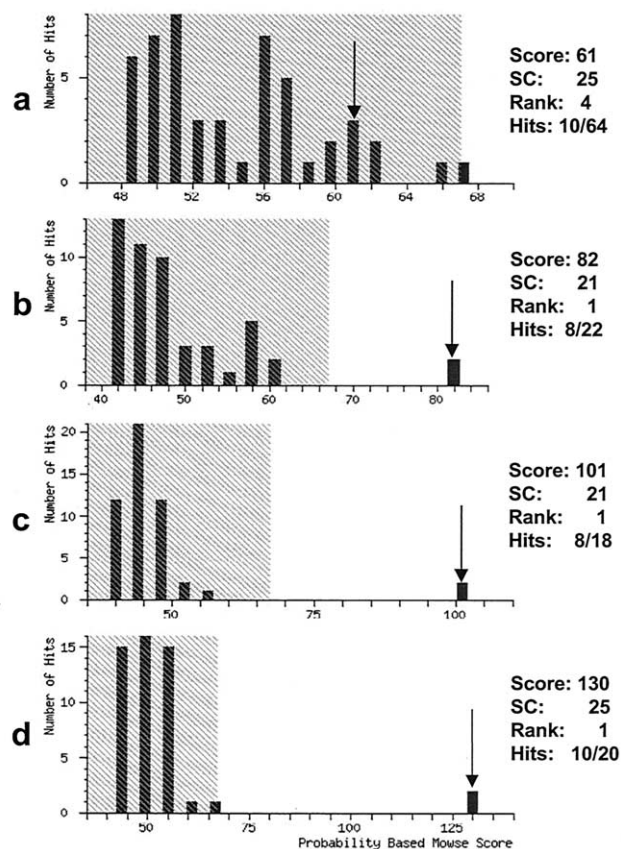
Figure 9. Example for improvement of spot identification using the iterative procedure combining MS-Screener analysis for removal of contaminants and cluster analysis for removal of peptide masses from contaminating neighbor spots. (**a**) Initial search with complete peak list; the score (61) of the finally identified protein (response regulator HP0703) does not reach the significance level; (**b**) search with the modified peak list after the first iteration excludes determined contaminants, and yields the score 82, clearly differentiating the identified protein from other candidates; (**c**) the search with the further modified peak list with additional excluded neighbor spot peptide masses yields a score 101; (**d**) the search with the final peak list, after reintroduction of two of the "contaminant" peaks that also fit the main component yields a final score of 130 and the sequence coverage of 25%. Three main peaks fit after removal of contaminants with two orthologous response regulators of two *H. pylori* strains (J99 and 26695), fulfilling our criteria for identification. HP0703 was the one of the investigated strain 26695.

must be eliminated by sequencing of the peptide. Additional help for the decision may be obtained by the ChemScore values [19], which can be obtained in MS Fit for each peptide. In nearly all of our analyses there are after this procedure still some mass peaks remaining, which may be caused by unknown posttranslational modifications, further contaminants, and unspecific cleavages.

As an application of the iterative procedure the analysis of spot number 2_175 (Figure 9) resulted in the identification of a main component (response regulator HP0703), which was not achievable from the raw PMF data. The scoring factor 61 of the finally identified protein did not distinguish between several protein

candidates (Figure 9a). By removing the contaminants the scoring factor increased to 82 with a clear distance to other candidates, unfortunately accompanied by a decrease of SC from 25 to 21% (Figure 9b). After removing neighbor spot peptide masses the scoring factor further increased to 101 (Figure 9c). The third iteration step, the reevaluation of the contaminants, showed that two of the "contaminant" peaks also fit the main component. These two masses, 763.41 and 1045.52 were first assigned to ions of DHB and trypsin with theoretical masses of 763.40 and 1045.59, respectively. In the last iteration step the assignments of these two masses were checked again with further criteria: Mass accuracy and probability of other peaks of the proposed contaminant. Peak 763.40 (intervalized value) occurred only in about 5% of all DHB spectra (Figure 4a), other more prominent DHB peaks such as 530.5, 558.5, 586.6, and others are missing in the spectrum. Therefore, it is more probable that the experimental peak of 763.41 belongs to the peptide of HP0703 with the theoretical mass of 763.43. In the case of peak 1045.52 (EITRLGDNK) the Δ ppm to the trypsin peptide is +66, to the HP0703 peptide it is −42. The Δ ppm values of the flanking and fitting HP0703 peptides are −45 (ELLGVVER) and −28 (VEEILPIAEIK), clearly suggesting that the mass of 1045.52 belongs to the HP0703 peptide and not to the trypsin peptide. It is clear that these are probability considerations and confirmation by MS/MS is necessary. However, the iterative procedure dramatically improves our ability to decide which of the masses should be analyzed by MS/MS. Reintroducing these two peaks resulted in a score of 130 and a SC of 25% for the main component (Figure 9d). After this analysis 10 mass peaks from 64 remain to be assigned. These ten masses fulfill the half-decimal place rule and have to be further analyzed by MS/MS or searched for metastable products as suggested recently [38]. The iterative data analysis was applied for seven other primarily not identified spots (Table 3). For all eight spots the MASCOT score factors were clearly increased during iterations and the three main peaks of each spectrum matched with the identified protein, fulfilling our criteria for identification. Nearly all of the matched peaks had ChemScore values [19] higher than ten confirming identification. Two spots, 4_216 and 4_240, gave poor MASCOT scores which resulted from the low number of peaks, but the ChemScore values of the three main peaks were convincing (4_216: 1204.68, 1553.78 = 100 and 1968.96 = 10; 4_240: 692.37, 862.47 = 100 and 1091.58 = 10).

The proposed procedure combines the knowledge accumulated from a large data set and tries to understand the protein composition of 2-DE spots. The data set of this study represents a kind of worst-case scenario, because preparative gels with a poorer resolution as compared to thin analytical gels were used. In addition, the view on exactly excised spots revealed much less neighbor spot contaminants in contrast to spot-sized but not spot-focused pixels. Nevertheless, a

Table 3. Confirmation of eight candidates by iterative data analysis. The MASCOT searches were restricted using the Proteobacteria (NCBInr 20030311) database with variable modifications; N-Acetyl (Protein), oxidation (M), propionamide (C), and a peptide mass tolerance of ±100 ppm

Spot number	Peak number			Main component	Final SC [%]	Ranking[a]		MASCOT scoring factor[a]		Matched 3 main peaks*
	Total	Contaminants[c]	Neighbor spot contaminants[d]			First	Final	First	Final	
2_175	64	42	3	HP0703	25	4	1	61	230	yes
2_214	24	8	0	HP0396	22	1	1	129	161	yes
2_300	52	32	3	HP0269	18	—[b]	2	—	74	yes
2_388	82	41	3	HP0116	24	1	1	82	133	yes
4_032	40	20	2	HP1014	26	3	1	48	67	yes
4_216	38	15	5	HP1563	29	—	1	—	51	yes
4_240	39	21	0	HP0175	13	—	4	—	48	yes
4_331	57	27	9	HP0802	40	23	1	41	60	yes

*Peptides of the 3 main peaks had ChemScore [19] values > 10.
[a]First ranking and score factors were obtained by searches with the complete mass lists and the final values resulted after the last iteration step.
[b][—] No *H. pylori* candidate under the first 50 reported hits.
[c]Contaminants were determined with MS-Screener.
[d]Neighbor spots contaminants were detected by cluster analysis and represent peptide masses which don't reach an SC of 30%. All of the eight spots contain only one main component.

procedure to analyze the complete composition of a spot, if possible with quantitative relations between the different components, will be important to evaluate the identification result with respect to the biological effect under investigation.

Primary MS data have to be analyzed by the half-decimal place rule, internally calibrated by polymers [39], and analyzed by several iterations to understand the composition of a spot. Implementation of the three iteration steps described in this study in automated identification starting with PMFs will give a more reliable view on spots of 2-DE gels.

Conclusions

A first search of PMFs in sequence databases normally results in many unassigned masses. These masses are the result of contaminants, masses of proteins from neighbor spots, additional main components of the spot, and post-translational modifications. We were able to remove contaminants using MS-Screener analysis. Neighbor spot masses were recognized by cluster analysis. A third iteration where contaminants were reevaluated increased the score of the main components. It should be stressed again that for each data set the contribution of contaminants and neighbor spots have to be determined. Even if the entire experiment is run under completely standardized and sterile conditions, masses of trypsin autolysis products, dye molecules, matrix, and neighbor spot peptides have to be eliminated. Increasing the number of analyzed spots using this iterative approach will result in an increased quality of the identification of proteins and protein species within each single spot. This procedure will help to understand the protein composition of each spot and correlation between different spots.

Acknowledgments

References

1. Klose, J. Protein Mapping by Combined Isoelectric Focusing and Electrophoresis in Mouse Tissues. A Novel Approach to Testing for Induced Point Mutations in Mammals. *Humangenetik* **1975,** *26,* 231–243.
2. O'Farrell, P. H. High-Resolution Two-Dimensional Electrophoresis of Proteins. *J. Biol. Chem.* **1975,** *250,* 4007–4021.
3. Aebersold, R. H.; Teplow, D. B.; Hood, L. E.; Kent, S. B. Electroblotting onto Activated Glass: High Efficiency Preparation of Proteins from Analytical Sodium Dodecyl Sulfate Polyacrylamide Gels for Direct Sequence Analysis. *J. Biol. Chem.* **1986,** *261,* 4229–4238.
4. Vandekerckhove, J.; Bauw, G.; Puype, M.; VanDamme, J.; VanMontagu, M. Protein Blotting on Polybrene-Coated Glass-Fiber Sheets. A Basis of Acid Hydrolysis and Gas-Phase Sequencing of Picomole Quantities of Protein Previously Separated on SDS-Polyacrylamide Gel. *Eur. J. Biochem.* **1985,** *152,* 9–19.
5. Fenn, J. B. Mann M.;; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989,** *246,* 64–71.
6. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. Protein and Polymer Analyses up to *m/z* 100,000 by Laser Ionization Time-of-Flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 151–153.
7. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60,* 2299–301.
8. Cleveland, D. W.; Fischer, S. G.; Kirschner, M. W.; Laemmli, U. K. Peptide Mapping by Limited Proteolysis in Sodium Dodecyl Sulfate and Analysis by Gel Electrophoresis. *J. Biol. Chem.* **1977,** *252,* 1102–1106.

9. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90,* 5011–5015.

10. Pappin, D.; Hojrup, P.; Bleasby, A. J. Rapid Identification of Proteins by Peptide Mass Fingerprinting. *Curr. Biol.* **1993,** *3,* 327–332.

11. Langen, H.; Takacs, B.; Evers, S.; Berndt, P.; Lahm, H. W.; Wipf, B.; Gray, C.; Fountoulakis, M. Two-Dimensional Map of the Proteome of. *Haemophilus influenzae. Electrophoresis* **2000,** *21,* 411–429.

12. Jungblut, P. R.; Schaible, U. E.; Mollenkopf, H.-J.; Zimny-Arndt, U.; Raupach, B.; Mattow, J.; Halada, P.; Lamer, S.; Kaufmann, S. H. E. Comparative Proteome Analysis of *Mycobacterium tuberculosis*and *Mycobacterium bovis* BCG Strains: Towards Functional Genomics of Microbial Pathogens. *Mol. Microbiol.* **1999,** *33,* 1103–1117.

13. Jungblut, P. R.; Bumann, D.; Haas, G.; Zimny-Arndt, U.; Holland, P.; Lamer, S.; Siejak, F.; Aebischer, A.; Meyer, T. F. Comparative Proteome Analysis of. *Helicobacter pylori. Mol. Microbiol.* **2000,** *36,* 710–725.

14. Ueberle, B.; Frank, R.; Herrmann, R. The Proteome of the Bacterium *Mycoplasma pneumoniae*: Comparing Open Reading Frames to Identified Gene Products. *Proteomics* **2002,** *2,* 754–764.

15. Lamer, S.; Jungblut, P. R. Matrix-Assisted Laser Desorption-Ionization Mass Spectrometry Peptide Mass Fingerprinting for Proteome Analysis: Identification Efficiency after On-Blot or In-Gel Digestion with and without Desalting Procedures. *J. Chromatogr. B* **2001,** *752,* 311–322.

16. Mattow, J.; Jungblut, P. R.; Schaible, U.; Mollenkopf, H.-J.; Lamer, S.; Zimny-Arndt, U.; Hagens, K.; Müller, E.-C.; Kaufmann, S. H. E. Identification of Proteins from *Mycobacterium tuberculosis* Missing in Attenuated *Mycobacterium bovis* BCG Strains. *Electrophoresis* **2001,** *22,* 2936–2946.

17. Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Jungblut, P. R.; Rudel, T. Analysis of Missed Cleavage Sites, Tryptophan Oxidation and N-Terminal Pyroglutamylation after In-Gel Tryptic Digestion. *Rapid Commun. Mass Spectrom.* **2000,** *14,* 496–502.

18. Krause, E.; Wenschuh, H.; Jungblut, P. R. The Dominance of Arginine-Containing Peptides in MALDI-Derived Tryptic Mass Fingerprints of Proteins. *Anal. Chem.* **1999,** *71,* 4160–4165.

19. Parker, C. P. Scoring Methods in Peptide Mass Fingerprinting: Chemscore, and the ChemApplex Program. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 22–39.

20. Jungblut, P.; Otto, A.; Zeindl-Eberhart, E.; Pleissner, K.-P.; Knecht, M.; Regitz-Zagrosek, V.; Fleck, E.; Wittmann-Liebold, B. Protein Composition of the Human Heart: The Construction of a Myocardial Two-Dimensional Electrophoresis Database. *Electrophoresis* **1994,** *15,* 685–707.

21. Klose, J.; Kobalz, U. Two-Dimensional Electrophoresis of Proteins: An Updated Protocol and Implications for a Functional Analysis of the Genome. *Electrophoresis* **1995,** *16,* 1034–1059.

22. Otto, A.; Thiede, B.; Müller, E.-C.; Scheler, C.; Wittmann-Liebold, B.; Jungblut, P. Identification of Human Myocardial Proteins Separated by Two-Dimensional Electrophoresis Using an Effective Sample Preparation for Mass Spectrometry. *Electrophoresis* **1996,** *17,* 1643–1650.

23. Schmidt, F. Lueking A.; Nordhoff, E.; Gobom, J.; Klose, J.; Seitz, H.; Egelhofer, V.; Eickhoff, H.; Lehrach, H.; Cahill, D. J. Generation of Minimal Protein Identifiers of Proteins from Two-dimensional Gels and Recombinant Proteins. *Electrophoresis* **2002,** *23,* 621–625.

24. Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data.* Prentice-Hall: Englewood Cliffs, NJ, 1988.

25. Mann, M. Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Applied Topics, Atlanta, GA, 1995; pp 639–640.

26. Karty, J. A.; Ireland, M. M. E.; Brun, Y. V.; Reilly, J. P. Artifacts and Unassigned Masses Encountered in Peptide Mass Mapping. *J. Chromatogr.* **2002,** *782,* 363–383.

27. Harris, W. A.; Janecki, D. J.; Reilly, J. P. Use of Matrix Cluster and Trypsin Autolysis Fragments as Mass Calibrants in Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2002,** *16,* 1714–1722.

28. Binz, P.-A.; Muller, M.; Walther, D.; Bienvenut, W. V.; Gras, R.; Hoogland, C.; Bouchet, G.; Gasteiger, E.; Fabbretti, R.; Gay, S. A. Molecular Scanner to Automate Proteomic Research and to Display Proteome Images. *Anal. Chem.* **1999,** *71,* 4981–4988.

29. Müller, M.; Gras, R.; Binz, P.-A.; Hochstrasser, D. F.; Appel, R. D. Molecular Scanner Experiment with Human Plasma: Improving Protein Identification by Using Intensity Distributions of Matching Peptide Masses. *Proteomics* **2002,** *2,* 1413–1425.

30. Scheler, C.; Lamer, S.; Pan, Z.; Li, X-P.; Salnikow, J.; Jungblut, P. Peptide Mass Fingerprint SC from Differently Stained Proteins on 2-DE Patterns by MALDI-MS. *Electrophoresis* **1998,** *19,* 918–927.

31. Lopez, M. F.; Berggren, K.; Chernokalskaya, E.; Lazarev, A.; Robinson, M.; Patton, W. F. A Comparison of Silver Stain and Sypro Ruby Protein Gel Stain with Respect to Protein Detection in Two-Dimensional Gels and Identification by Peptide Mass Profiling. *Electrophoresis* **2000,** *21,* 3673–3683.

32. Rabilloud, T.; Strub, J.-M.; Luche, S.; van Dorsselaer, A.; Lunardi, J. A Comparison between Sypro Ruby and Ruthenium II tris(Bathophenantroline Disulfonate) as Fluorescent Stains for Protein Detection in Gels. *Proteomics* **2001,** *1,* 699–704.

33. Mujumdar, R. B.; Ernst, L. A.; Mujumdar, S. R.; Waggoner, A. S. Cyanine Dye Labeling Reagents Containing Isothiocyanate Groups. *Cytometry* **1989,** *10,* 11–19.

34. Uenlue, M.; Morgan, M. E.; Minden, J. S. Difference Gel Electrophoresis: A Single Method for Detecting Changes in Protein Extracts. *Electrophoresis* **1997,** *18,* 2071–2077.

35. Scheler, C.; Müller, E.-C.; Stahl, J.; Müller-Werdan, U.; Salnikow, J.; Jungblut, P. Identification and Characterization of Heat Shock Protein 27 Protein Species in Human Myocardial 2-DE Patterns. *Electrophoresis* **1997,** *18,* 2823–2831.

36. Mattow, J.; Jungblut, P. R.; Müller, E.-C.; Kaufmann, S. T. H. Identification of Acidic, Low Molecular Mass Proteins of Mycobacterium tuberculosis Strain H37Rv by MALDI- and ESI-Mass Spectrometry. *Proteomics* **2001,** *1,* 494–507.

37. Aksu, S.; Scheler, C.; Focks, N.; Leenders, F.; Theuring, F.; Salnikow, J.; Jungblut, P. R. An Iterative Calibration Method With Prediction of Posttranslational Modifications for the Construction of a Two-Dimensional Electrophoresis Database of Mouse Mammary Gland Proteins. *Proteomics* **2002,** *2,* 1452–1463.

38. Wirth, U.; Müller, D.; Schindler, P.; Lange, J.; Van Oostrum, J. Post-Translational Modification Detection Using Metastable Ions in Reflector Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry. *Proteomics* **2002,** *2,* 1445–1451.

39. Gobom, J.; Mueller, M.; Egelhofer, V.; Theiss, D.; Lehrach, H.; Nordhoff, E. A Calibration Method that Simplifies and Improves Accurate Determination of Peptide Molecular Masses by MALDI-TOF MS. *Anal. Chem.* **2002,** *74,* 3915–3923.