Analysis of Automatically Generated Peptide Mass Fingerprints of Cellular Proteins and Antigens from *Helicobacter pylori* 26695 Separated by Two-dimensional Electrophoresis*

Alexander Krah‡§¶, Frank Schmidt‡, Dörte Becher∥, Monika Schmid‡, Dirk Albrecht∥, Axel Rack‡, Knut Büttner∥, and Peter R. Jungblut‡

*Helicobacter pylori* is a causative agent of severe diseases of the gastric tract ranging from chronic gastritis to gastric cancer. Cellular proteins of *H. pylori* were separated by high resolution two-dimensional gel electrophoresis. A dataset of 384 spots was automatically picked, digested, spotted, and analyzed by matrix-assisted laser desorption ionization mass spectrometry peptide mass fingerprint in triple replicates. This procedure resulted in 960 evaluable mass spectra. Using a new version of our data analysis software MS-Screener we improved identification and tested reliability of automatically generated data by comparing with manually produced data. Antigenic proteins from *H. pylori* are candidates for vaccines and diagnostic tests. Previous immunoproteomics studies of our group revealed antigen candidates, and 24 of them were now closely analyzed using the MS-Screener software. Only in three spots minor components were found that may have influenced their antigenicities. These findings affirm the value of immunoproteomics as a hypothesis-free approach. Additionally, the protein species distribution of the known antigen GroEL was investigated, dimers of the protein alkyl hydroperoxide reductase were found, and the fragmentation of γ-glutamyltranspeptidase was demonstrated. *Molecular & Cellular Proteomics* 2:1271–1283, 2003.

---

1. The abbreviations used are: 2-DE, two-dimensional electrophoresis; MW, molecular weight; PMF, peptide mass fingerprint; S/N, signal-to-noise ratio; MALDI, matrix-assisted laser desorption ionization; MS, mass spectrometry; TOF, time-of-flight; TOF/TOF, tandem time-of-flight; ORF, open reading frame; CBB, Coomassie Brilliant Blue; CHAPS, 3-(3-cholamidopropyl)dimethylammonio)-1-propanesulfonic acid.

From the ‡Core Facility for Protein Analysis and §Department of Molecular Biology, Max Planck Institute for Infection Biology, 10117 Berlin, and the ∥Institute for Microbiology, Ernst Moritz Arndt University Greifswald, 17487 Greifswald, Germany

Received, August 14, 2003, and in revised form, September 29, 2003

Published, MCP Papers in Press, September 29, 2003, DOI 10.1074/mcp.M300077-MCP200

© 2003 by The American Society for Biochemistry and Molecular Biology, Inc.

This paper is available on line at http://www.mcponline.org
Experimental Procedures

H. pylori Cell Culture and Lysis—Bacteria were grown on agar plates containing 10 μg/ml vancomycin. After 3 days single clones were resuspended in 1 ml of brain heart infusion medium containing 10% fetal calf serum, and 20 μl of this suspension were grown for 2 days on vancomycin-containing agar plates under microaerobic conditions (5% O2, 10% CO2, 85% N2) at 37 °C. Cells were transferred into 50 ml of cold phosphate-buffered saline containing Complete protease inhibitors (Roche Applied Science). After centrifugation at 100,000 × g for 10 min, one wash step in 10 ml of phosphate-buffered saline containing protease inhibitor followed. The pellet of bacteria was diluted with a half-volume of distilled water and lysed by addition of urea, CHAPS, Servalyte, pI 2–30-cm high resolution gel system with a resolution power of up to 5,000 spots (1, 17). For the first dimension an ampholyte mix of pI 2–70, and the second dimension ranged from 5–300 kDa. For preparative gels 250 μg of protein were loaded, and the gels were stained with Coomassie Brilliant Blue (CBB) G-250 for 5 days (18).

Two-dimensional Electrophoresis—H. pylori lysate proteins were separated using a 23 × 30-cm high resolution gel system with a resolution power of up to 5,000 spots (1, 17). For the first dimension an ampholyte mix of pl 2–11 was used, no alkylation was performed, and the second dimension ranged from 5–130 kDa. For preparative gels 250 μg of protein were loaded, and the gels were stained with Coomassie Brilliant Blue (CBB) G-250 for 5 days (18).

In-gel Digest—In a large section of the 2-DE gel (20–80 kDa, whole pl range) 384 different spots with staining intensities ranging from very weak to very high were excised in triple replicates. For this purpose a spot cutter (Proteome Works™ Bio-Rad) with a picker head of 2-mm diameter was used. Cut spots were transferred into 96-well microtiter plates. The tryptic digest with subsequent spotting on a matrix-assisted laser desorption ionization target was carried out automatically with the Ettan™ spot-handling work station (Amersham Biosciences) using the following protocol. The gel pieces were washed twice with 100 μl of a solution of 50% CH3CN and 50% 50 mM NH4HCO3 for 30 min and washed once with 100 μl of 75% CH3CN for 10 min. After drying at 37 °C for 17 min 10 μl of trypsin solution containing 20 ng/μl trypsin (Promega, Madison, WI) was added and incubated at 37 °C for 120 min. For extraction, gel pieces were covered with 60 μl of 0.1% trifluoroacetic acid in 50% CH3CN and incubated for 30 min at 40 °C. The peptide containing supernatant was transferred into a new microtiter plate, and the extraction was repeated with 40 μl of the same solution. The supernatants were dried at 40 °C for 220 min. The dry residue was dissolved in 3 μl of 0.5% trifluoroacetic acid in 50% CH3CN, and 0.4 μl of this solution was directly spotted onto the matrix-assisted laser desorption ionization target. Then 0.4 μl of a saturated α-cyano-4-hydroxyphenylacetic acid solution in 70% CH3CN was added and mixed with the sample by aspirating the mixture five times. The samples were allowed to dry on the target for 10–15 min before measurement in MALDI-TOF.

MALDI-TOF Mass Spectrometry—The MALDI-TOF measurement was carried out on the 4700 Proteomics Analyzer (Applied Biosystems, Foster City, CA). This instrument is designed for high throughput measurement, being automatically able to measure the samples, calibrate the spectra, and process the data using the 4700 Explorer™ software. The spectra were recorded in a mass range from 900 to 3,700 Da with a focus mass of 2,000 Da. For one main spectrum 20 subspectra with 100 shots/subspectrum were accumulated using a random search pattern. If the autolytic fragment of trypsin with the monoisotopic (M + H)+ m/z at 2,211.104 reached a signal-to-noise ratio (S/N) of at least 10, an internal calibration was automatically performed as one-point calibration using this peak. If the automatic mode failed, manual calibration was applied. After calibration peak lists were created by using the "peak-to-mascot" script of the 4700 Explorer™ software. Settings were a mass range from 900 to 3,500 Da, a peak density of 50 peaks/200 Da, a minimal area of 0, and a maximum of 200 peaks/spot. Three different peak lists were created for an S/N ratio of 5, 7, and 10, respectively. For confirmation of selected peaks MALDI-TOF/TOF spectra were recorded manually.

Database Searches—Identification of spots was done via batch mode using the Mascot protein identification system (Matrix Science Ltd., London, UK) in-house applying the recent H. pylori 26695 protein database downloaded from The Institute for Genomic Research (TIGR, www.tigr.org/). Optimal search parameters were 30 ppm peptide mass tolerance, fixed oxidation of methionine, and 1 missed trypsin cleavage. The criterion for reliable identification was a significant Mascot score >45 (p < 0.05) (19).

Data Analysis with MS-Screener—To realize an iterative data analysis (18) for large datasets we have developed a new MS-Screener version using the Java 2 standard edition 1.4.1 software development kit (J2SE1.4.1 SDK, java.sun.com/). This Java tool consists of 126 different program classes and was integrated in a user-friendly graphical user interface (GUI). To integrate a plot view the JFreeChart class library, version 0.9.8, was applied (www.jfree.org/jfreechart/index.html). The software runs under LINUX, Solaris, and Microsoft Windows and comprises a setup function for all operating systems. MS-Screener has the ability to import different ASCII file types like .pkm (GRAMS), .pkt, .txt (Data Explorer), and .dta (SEQUEST). Contaminant searches, calculation of half-decimal places, elimination of contaminants, screening of common masses, their rankings, and the generation of matrices to realize hierarchical agglomerative cluster analyses using R (www.r-project.org) can now easily be calculated in one work set. To find common contaminants in the complete dataset a mass tolerance (interval width) of 30 ppm and a threshold of 5% were applied. Masses that exceeded this threshold were eliminated from the peak lists. To calculate the half-decimal place rule an absolute standard deviation of 0.12 Da was applied, and outlier masses were marked and extracted in a separated table. Another function of the MS-Screener allowed us to generate binary or non-binary interval matrices, which include all intensity values of peak intervals as zero/one or real intensity counts, respectively. In the present study we used an interval width of 30 ppm, and about 1,600 intervals were calculated based on 384 spectra for one gel. Using these matrices
hierarchical agglomerative cluster analyses were performed using the statistical programming environment R (www.r-project.org).

RESULTS

Automation of the Identification Process—H. pylori lysate proteins were separated by high resolution 2-DE. A number of spots were selected from a large section of the gel ranging from 20–80 kDa and over the whole pI range (4.5–10). Staining intensities went from very low to very high. Triple replicates of 384 spots were automatically cut, digested, and measured by MALDI-MS. Automatic processing yielded 88% of spectra calibrated, and an additional 3% of spectra were calibrated manually. This procedure resulted in a dataset of 960 evaluable spectra; one such spectrum is shown in Fig. 1.

An effective identification of a large number of spectra can be done quickly using an automated system. However, search results of spectra with divergent quality are strongly influenced by the parameters applied. Therefore we tested different parameters to find out whether an optimum for our dataset exists. For this purpose nearly 6,000 database searches were performed with peak lists created with S/N for peak detection of 5, 7, and 10. Additionally, different fixed and variable modifications as well as peptide mass tolerances were used. The following parameters produced the highest identification rates: a peak detection with S/N of 7, fixed methionine oxidations, peptide mass tolerance of 30 ppm, and a maximum of one missed trypsin cleavage. Using this set of parameters 547 spectra (equalling 57%) were automatically identified. Due to the triple replicate dataset 75% of different spots were identified at least once.

The automated recording and one-point calibration (using the autolytic trypsin peptide m/z 2,211.104) of mass spectra yielded peak masses that allowed database searches with an optimal mass tolerance of 30 ppm. The error distribution showed a slight slope to negative errors for smaller masses. This could have been avoided by applying a two-point calibration. However, many spectra did not contain a second sufficiently intense autolytic peak of trypsin (e.g. m/z 1,045.6). The described error distribution remained stable over all measured spectra.

An interesting finding was that possible doubly charged peaks occurred in rare cases of PMFs. Such peaks were conspicuous with regard to the half-decimal place rule, and isotopic patterns showed peak distances of m/z/H110050.5. For example in spot 433d (HP0410) the peptide containing amino acids 31–49 (QQHNTGESVELHFHYPIK) was found with high intensity as a single-charged peptide with m/z 2,278.1 and as a doubly charged peptide with m/z 1,139.5. This peptide contains three histidines as potential additional proton acceptors.

To evaluate the reliability of the automated identifications we searched for contradictions in identifications of replicate spots. Here only nine spots (3%) were found to be differently...
identified with significant scores within the three replicates. When looking closer into these contradictions three spots were found to be imprecisely picked in a densely "populated" area. Two spots contained minor components caused by smearing from neighboring spots that were erroneously identified as major components. Furthermore, three spots contained mixtures of proteins where either one of them was identified as the major component. Only one spectrum was erroneously assigned to a protein.

Another way to evaluate the automated procedure was to compare identification data with manually produced data using a Voyager Elite (Applied Biosystems) (20). Sequence coverages and Mascot scores of 28 arbitrarily chosen spots with medium to low staining intensities were compared (Table I). For this purpose automatically acquired (S/N 7, contaminants removed) and manually measured peak lists (manual peak detection, contaminants also removed (16)) were compared using similar database searches with the Mascot protein identification system. Search parameters were similar apart from the peptide mass tolerances (automatic, 30 ppm; manual, 100 ppm) and possible methionine oxidations (automatic, fixed; manual, variable). For medium staining intensities both methods obtained comparable results. For the more interesting weakly stained spots, however, most spots showed more matched peaks, higher sequence coverages, and higher Mascot scores using the manual procedure.

Improvement of Identification, the Search for Spot Contaminants, and Additional Spot Components

To improve identification and detect minor spot components in large datasets we developed the iterative data analysis software MS-Screener (16). The new version integrates tools for contaminant search and removal, calculation and plotting of the decimal places of masses, and calculation of interval matrices to perform clusterings in R (Fig. 2). Here a threshold of 5% (mass

| Spot no. | TopSpot ID | TIGR locus | Matched peaks/all peaks | Sequence coverages | Mascot scores |
|----------|------------|------------|--------------------------|-------------------|--------------|
|          |            |            | automatic | manual | automatic | manual | automatic | manual | automatic | manual |
| medium staining intensities |
| 359 | 3_148 | HP0958 | 10/26 | 4/6 | 37 | 20 | 135 | 64 |
| 363 | 3_170 | HP0067 | 6/19 | 10/21 | 26 | 36 | 77 | 101 |
| 384 | 4_313 | HP1562 | 13/23 | 10/13 | 38 | 27 | 189 | 118 |
| 420 | 4_112 | HP0139 | 5/7 | 10/32 | 19 | 33 | 89 | 93 |
| 432 | 4_340 | HP0990 | 3/25 | 7/17 | 10 | 21 | 25 | 74 |
| 436 | 4_321 | HP0073 | 11/20 | 15/26 | 43 | 54 | 176 | 184 |
| 440 | 4_326 | HP0073 | 10/19 | 12/18* | 39 | 33 | 155 | 129 |
| 458 | 4_277 | HP0485 | 12/20 | 12/15 | 39 | 38 | 183 | 171 |
| 483 | 4_097 | HP0630 | 6/12 | 8/19 | 33 | 40 | 101 | 115 |
| average values | | | 8.4/19.0 | 9.8/18.6 | 31.6 | 33.6 | 125.6 | 117.7 |
| low staining intensities |
| 118 | 1_205 | HP0123 | 4/8 | 14/21 | 7 | 28 | 48 | 155 |
| 132 | 2_018 | HP0696 | 9/16 | 27/36 | 12 | 35 | 98 | 279 |
| 209 | 1_050 | HP1293 | 9/20 | 14/26 | 39 | 49 | 127 | 137 |
| 219 | 1_160 | HP0154 | 3/3 | 12/35 | 14 | 46 | 54 | 114 |
| 236 | 1_227 | HP0197 | 5/11 | 9/22 | 17 | 29 | 68 | 99 |
| 239 | 1_440 | HP0589 | 5/9 | 8/25* | 12 | 30 | 69 | 67 |
| 302 | 1_120 | HP0077 | 2/2 | 10/30 | 8 | 38 | 40 | 110 |
| 318 | 2_081 | HP1164 | 5/10 | 7/10 | 22 | 28 | 76 | 104 |
| 360 | 3_189 | HP1588 | 3/6 | 7/15 | 17 | 33 | 50 | 94 |
| 364 | 3_150 | HP0176 | 8/6 | 9/11 | 29 | 40 | 125 | 147 |
| 365 | 3_177 | HP0176 | 4/11 | 4/6 | 22 | 13 | 60 | 59 |
| 367 | 4_026 | HP1299 | 6/15 | 7/26 | 26 | 37 | 87 | 80 |
| 388 | 4_167 | HP1024 | 9/14 | 11/20 | 41 | 48 | 148 | 150 |
| 393 | 4_203 | HP0073 | 6/9 | 14/25 | 23 | 50 | 102 | 168 |
| 418 | 4_084 | HP0069 | 9/21* | 11/40 | 43 | 51 | 121 | 86 |
| 425 | 4_134 | HP1102 | 9/15 | 12/24 | 44 | 53 | 151 | 167 |
| 473 | 4_230 | HP0954 | 3/3 | 7/13 | 15 | 30 | 63 | 88 |
| 480 | 4_127 | HP1533 | 8/11 | 6/8 | 37 | 27 | 144 | 101 |
| 487 | 3_084 | HP0794 | 2/7 | 9/23* | 23 | 44 | 37 | 83 |
| average values | | | 5.7/10.4 | 10.4/21.9 | 23.7 | 37.3 | 87.8 | 120.4 |
occurred in ≥48 of 960 evaluable spectra) was used to define contaminant peaks in the dataset (Fig. 3 and Table II). Sixty-one masses were found to be contaminants; 12 of these were trypsin autolytic peaks, and 12 were matrix cluster peaks (α-cyano-4-hydroxycinnamic acid, Na⁺, and K⁺ clusters were outliers of the half-decimal place rule; cluster masses were calculated as described by Keller and Li (21)). Four peaks were unknown outliers of the half-decimal place rule, three peaks belong to the most intense peaks of GroEL, seven were erroneously labeled isotope peaks because of low peak intensities, and the remaining peaks were unknown. It is important to notice that no keratin peaks were found. After removing the 5% most frequently occurring masses in the dataset the identification of spots was improved by 3% to 78% of all spots to be identified at least in one gel (using optimal parameters). A list of all identified spots is found in the supplemental table. As expected after removal of contaminants spot identification was improved for most spots except for very intensely stained ones; three masses of GroEL were misleadingly defined as contaminants (Table II). Even though this most widely spread protein was identified in 15 different spots only (3.9% of spots) these three peptide masses occurred in more than 6% of the spectra.

**Protein Species in the 2-DE Gels**—Expression products of many ORFs appear modified as different protein species in the form of several spots within one 2-DE gel. Examples of this phenomenon are the following identified proteins: translation elongation factor EF-Tu (HP1205, four spots), catalase (HP0875, five spots), alkyl hydroperoxide reductase (HP1563, eight spots), urease α subunit (HP0073, nine spots), and chaperone and heat shock protein GroEL (HP0010, 15 spots). On average we found 1.6 different protein species/ORF in our dataset. With 15 different spots GroEL occurred most frequently (Fig. 4, left). Interestingly, three groups of GroEL spots occurred in the 2-DE gels: a main spot group (five spots), one group with lower MW and more acidic pI (two spots), and one group with lower MW and more basic pI (six spots). Evidence was found that the second group is N-terminally truncated.
because one peptide found in the main spot (amino acids 13–20) was not found in the spectra of this spot group. The third group we assume to be C-terminally truncated GroEL protein species because seven peptides (comprising amino acids 425–522) were not found in the spectra of these spots. In-silico calculated and apparent (according to gel position) MWs and pIs were in good agreement (data not shown).

Even more than the 15 identified spots contained peptides from GroEL (Table III). When we searched all peak lists for the most intense peak from the GroEL main spot (m/z 1,595.9) 33 different spots were found. Using the more rigid criterion to contain at least three of the five most intensive peaks from GroEL (m/z 1,595.9, 947.5, 1,867.8, 1,488.7, and 1,401.6) 24 spots still were found. Three of these five peaks also belong to the contaminants list (1,595.9, 1,867.8, and 1,488.7), i.e. they were found in 7, 6, and 6% of the spectra, respectively. The other two were found in 4% of the spectra and therefore not considered as contaminants. GroEL peptides were distributed widely in the gels; however, the distribution differed from gel to gel. In gels A, B, and D we found 21, 12, and 13 spots that contained three of the five GroEL peptides, respectively (Table III).

The spot 312 was identified to be alkyl hydroperoxide reductase (HP1563); however, the apparent molecular weight of the spot position in the gel was about double the weight of the protein. We therefore assumed this spot to contain dimers of the protein. By comparing the PMFs (Fig. 5) of this spot with the proteins, main spot (413) differences in sequence coverage were seen. Both cysteine residues of the protein (amino acids 49 and 169) were not covered in spot 312; however, both were found in the main spot. First, amino acids 44–58/59 modified with propionamide (1,816.8), one missed cleavage as well as propionamide (1,973.0), and an additional methyl ester (1,987.0) were found in the PMF. Methyl ester formation is characteristic for our CBB G-250 staining in methanol and occurs frequently. Second, amino acids 155–174 with oxidized methionine and propionamide modification (2,383.1) were found, too. The sequences of peaks 1,973.0 and 2,383.1 were confirmed by MALDI-TOF/TOF. Because of significantly lower peak intensities in the PMF of spot 312 (staining intensity was also much lower), three of the peaks from spot 413 might not be detectable in spot 312; the peak m/z = 1,973.0, however, was more intense than the peak m/z = 1,649.8, which was found to be intense in both spectra. From this it follows that the cysteine 49-containing peptide was not seen in spot 312 and may therefore be involved in a disulfide bond formation to link the dimers. Because the peak of the second cysteine-containing peptide may be lost in the noise in spot 312, it cannot be distinguished whether homo- or heterodimers of the protein are formed.

A pair of spots with different molecular masses and the same identification (γ-glutamyltranspeptidase, Ggt, HP1118)
were spots 347 and 494. By comparing the sequences covered by the PMFs (Fig. 6) these spots appeared to be two fragments of the protein whose sequences were mutually exclusive. In-silico MW and pI calculation of the protein fragments with assumed cleavage at amino acid 370 resulted in similar values compared with the spot positions. Spot 347 is positioned at pI/MW coordinates 9.0/40.0, and amino acids 1–370 were calculated to have 9.5/39.8. For spot 494 we found 6.7/20.0 according to position, and 6.3/21.0 was calculated for amino acids 371–567. A spot containing the whole protein (theoretical mass of 61.2 kDa) was not found in the gels. Therefore, we assume that the entire H. pylori-glutamyl-transpeptidase content of the cell is processed into two subunits.

**Exploration of H. pylori Antigens**—Immunoproteomics is a method where 2-DE blots are incubated with antibodies, e.g., with human sera. Spots that are recognized can be identified using MALDI-MS. However, spots may contain minor components from other proteins or protein species that could have been recognized by highly specific antibodies instead of the main component. With the iterative procedure using MS-Screener and hierarchical clusterings we tested 24 H. pylori 26695 antigens known from previous studies to be differently recognized by patients suffering from diseases caused by...
H. pylori (gastritis, duodenal ulcer, and gastric carcinoma) or antigens known to be protective against H. pylori challenge in mice (Table IV) (10, 11, 22). Six antigens did not contain any reproducible peaks that were not assigned to the identified protein. With respect to the sensitivity of our method these spots can be assumed to be free of minor components. Nine spots contained peaks that could not be assigned to another protein; they may originate from modified peptides, unspecific cleavages, or unknown minor components. In the remaining nine spots peaks supposedly originating from a different protein were found (six of which were from a neighbor spot <1 cm apart). Apart from HP1533 and HP0380 all these minor components are known to be antigenic and were therefore further investigated as to whether they may have influenced the antigenicities of these spots. In immunoblots incubated with sera of H. pylori-infected patients we explored whether the spots with possible antigenic minor components were recognized concurrently with the main spots of these components. Evidence was found that antigen recognition of three spots might have been influenced by the minor component (spots 154, 278, and 279). For the other four spots no evidence for such an influence was found.

The antigenic protein GroEL (HP0010) was identified in 15 different spots in our dataset (see above and Fig. 4). Interestingly all of these spots were conjointly recognized by human sera from H. pylori-infected individuals. Searching the immunoblots for recognition of the 24 spots that contained three of the five most intense GroEL peaks (Table III) we found that all apart from three (spots 313, 314, and 372) were recognized conjointly by antibodies in human sera.

**Completion of the Proteome of H. pylori 26695—**Another aspect of this study was the continuation of the proteome exploration of H. pylori. Here we identified 298 spots (78% of the spots measured), which represent 183 different ORFs. Twenty-four of these ORFs have not been identified before as compared with the dataset of our group to be published (Table V). Among these, four spot identifications conflict with the manual results presumably because of spot-picking tolerances in densely spotted areas or because spots contain protein mixtures.

**DISCUSSION**

For this study an automatically generated dataset was used to compare identification results with our manual procedure, to exhaustively investigate protein distributions in 2-DE gels, and to affirm the immunoproteomics approach used to identify antigen candidates. Here we investigated a large dataset covering about two-thirds of visible spots (384 spots) of our CBB G-250-stained H. pylori 26695 2-DE gel. These spots were picked in triple replicates, digested, and measured automatically, and they resulted in a dataset of 960 evaluable spectra.

To take the most advantage of such a dataset it was shown to be helpful to optimize the peak detection and identification parameters. Not too many possible modifications should be used because Mascot scores will fall with increasing amounts of possible peptide masses. Even more important, however, is to take advantage of the recording of replicate datasets, which will further improve the rate of identification considerably. Performing searches in triple replicates, we were able to achieve an identification rate of 75% for spots that were finally identified in at least one gel.

A good approach to assess the reliability of automatic identifications is the search for contradictions in identifications of replicate spots. Such differences may be caused by spot-picking tolerances, incidental differences in the auto-

---

**Table III**

| Spot no. | GroEL peaks found |
|---------|-------------------|
|         | Gel A | Gel B | Gel C | Gel D |
| 154     | x     | o     | -     | x     |
| 157     | x     | x     | o     | -     |
| 158     | x     | -     | -     | -     |
| 159     | x     | -     | -     | -     |
| 160     | x     | -     | -     | -     |
| 161     | x     | -     | x     | -     |
| 162     | x     | -     | x     | -     |
| 163     | x     | x     | x     | x     |
| 164     | x     | x     | x     | x     |
| 165     | x     | x     | x     | x     |
| 206     | x     | -     | o     | -     |
| 218     | x     | -     | x     | -     |
| 223     | x     | o     | o     | o     |
| 224     | -     | x     | o     | o     |
| 225     | x     | x     | x     | x     |
| 226     | x     | x     | x     | x     |
| 227     | o     | x     | o     | o     |
| 231     | -     | x     | -     | -     |
| 232     | x     | x     | x     | x     |
| 233     | x     | x     | x     | x     |
| 234     | x     | x     | x     | x     |
| 313     | x     | o     | -     | -     |
| 314     | x     | o     | o     | -     |
| 372     | x     | -     | -     | -     |

Marked with *x* are spots in the corresponding gel(s) that contain at least three out of the five most intense GroEL peaks from the main spot (most intense peaks in descending order: 1,595.9, 947.5, 1,867.8, 1,488.7, and 1,401.6). Shaded in gray are spots that were unambiguously identified to contain GroEL as main component. The *o* means that this spot was identified to contain a protein different from GroEL (spot 154 was in gel A a mix of two proteins); others were not identified. The search was performed using MS-Screener with peak lists including contaminants.
matic procedure, or unsuited database search parameters. We found only 3% of spots to be inconsistently identified in the three replicate datasets. Many of these spots were positioned in densely spotted areas, and their inconsistent identification may therefore be more a problem of picking or of protein mixtures in spots than of erroneous database search.

**Fig. 5.** A protein and its dimer: a sector of a 2-DE gel, PMFs of two spots, and tandem MS spectra of two peaks. The *H. pylori* 2-DE gel was stained with CBB G-250 (upper left). The upper spot (312) contains dimers of the protein found in spot 413. Alkyl hydroperoxide reductase (HP1563) was identified in both spots with sequence coverages of 48 and 72%, respectively (PMFs, to the right of the gel). The two cysteine-containing peptides (amino acids 49 and 169) were not found in the dimer spot (triangles). One peptide covered the amino acids 44–58/59 modified with propionamide (1,816.9), one missed cleavage and propionamide (1,973.0), and an additional methyl ester (1,987.0). The other cysteine-containing peptide consists of amino acids 155–174 with oxidized methionine and propionamide modification (2,383.1). Due to the lower peak intensities in the upper spot several contaminant masses were found in the PMF (black squares). The cysteine-containing peptides 1,973.0 and 2,383.1 were confirmed by MALDI-TOF/TOF with Mascot scores indicating extensive homology (lower spectra). Fragments of the peptides are labeled in the spectra, and y and b series are also marked in the sequence. An asterisk indicates a loss of NH₃, a zero indicates a loss of H₂O, and peaks marked with int are internal fragments.
results. Spots laying side-by-side and containing different proteins will merge into one another even when the protein concentration in the merging zone is below the detection limit of the staining. Small variances in spot picking can in such cases coincidentally cause different identifications for one spot. The same is true for spots that contain mixtures of proteins with similar concentrations. Only one of 298 identifications was incorrect, which shows that identification was highly reliable. Additionally, the use of the exclusive identification criterion of a significant Mascot score of 45 (for use of The Institute for Genomic Research (TIGR) H. pylori 26695 database, $p < 0.05$) appeared to be trustworthy. It was not necessary to consider sequence coverages or number of matched peptides. The fact that only 3% of spots were inconsistently identified showed also the reproducibility of spot patterns in our 2-DE gels.

An important aspect of this study was the comparison of automatic and manual procedures of data generation and identification. We have chosen 28 exemplary spots, which were identified automatically as well as manually (Table I). By comparing these results it became evident that differences for medial stained spots were negligible, whereas differences between identification of faint spots were noticeable. After removal of contaminant peaks (discussed below) manually generated spectra of faint spots contained on average more peaks, and also more peaks were matched to the given protein. The same holds true for sequence coverages and Mascot scores. These results were probably caused by the fact that the manual procedure could be adapted for individual spots with low protein contents. It is quite evident that automatic procedures may not be adjusted to all the spots in a gel where protein contents differ by several orders of magnitude. Consequently, manual measurements and data analyses are still powerful means to investigate faint spots.

We have developed the software MS-Screener, which not only is able to improve identification but also can be used for...
exhaustive data analysis. The new user-friendly graphical user interface allows the import of data in the form of ASCII files, calculates and removes contaminants, calculates and plots half-decimal places, screens and ranks for certain masses in spectra, and enables the generation of intervalized peak intensity matrices for further statistical analyses (Fig. 2). This tool was successfully utilized to improve identification, to analyze protein distributions, and to find minor components especially in \textit{H. pylori} antigens as discussed below.

The removal of contaminants resulted in an improvement of the identification rate by 3\% to 78\% of spectra that were identified at least in one of the replicate gels. Sixty-one masses were found in 5\% of the 960 spectra and were therefore defined to be contaminants, i.e. peaks that were not specific for a certain spot (Table II and Fig. 3). Contaminant masses were assigned to be matrix clusters, trypsin autolytic products, or peptides from the most frequently found protein GroEL. Seven masses were erroneously labeled isotope peaks. For peaks with low intensity the peak-labeling algorithm of peak-to-Mascot picked the more intense second isotopic peak instead of the monoisotopic. In these cases the monoisotopic masses were found in other spectra and appear also in the contaminants list. Although the source of the other contaminant masses is unknown, no keratin peaks were found. In our previous study (16), in which in 480 manually acquired and analyzed PMFs 69 contaminant masses in the comparable mass range of 900–3,500 Da were found, 47 masses were assigned to keratins. In another recent study of 118 spectra (23), 71 contaminants in the range of 900–3,500 Da were found in \( \geq 5\% \) of spectra, and 53 of these were keratins. These results show that although a comparable number of contaminants were found the use of fully automated spot picking, digesting, and spotting can be highly efficient to avoid contaminations with keratin.

The fact that one spot does not contain one protein but rather one protein can be distributed in several spots in the form of different protein species is well illustrated by the heat shock protein GroEL. This most widely distributed protein in our dataset was identified in 15 different spots (Fig. 4, left). Within these spots evidence was found that two were N-terminally truncated and that six were C-terminally truncated. The reasons for the exact spot positions within these groups (modifications, differences in lengths of truncations, or conformational differences) were not figured out. A further nine spots were found to contain three of the five most intense peptide masses of GroEL (Table III) and may therefore most likely contain low amounts of GroEL. In six spots GroEL was a minor component because they were identified to contain a different protein, and in three spots (not identified) this protein may be a minor or major component. These findings raise the question as to whether minor components represent co-migrating proteins, e.g. by protein-protein interactions during

### Table V

| TIGR* locus | TIGR common name | Spot no. |
|-------------|-----------------|---------|
| HP0030      | Hypothetical protein | 152 |
| HP0075      | Urease protein | 174 |
| HP0132      | L-serine deaminase | 175 |
| HP0138      | Conserved hypothetical iron-sulfur protein (spot contains a mix of two proteins) | 259 |
| HP0183      | Serine hydroxymethyltransferase | 272 |
| HP0201      | Fatty acid/phospholipid synthesis protein | 368 |
| HP0354      | Deoxyxylulose-5-phosphate synthase, putative | 119 |
| HP0380      | Glutamate dehydrogenase | 266 |
| HP0401      | 3-Phosphoshikimate 1-carboxyvinyltransferase | 284 |
| HP0624      | Solute-binding signature and mitochondrial signature protein | 330 |
| HP0710      | Conserved hypothetical protein | 139 |
| HP0852      | Hypothetical protein | 370 |
| HP0859      | ADP-L-glycero-D-mannoheptose-6-epimerase | 378 |
| HP0886      | Cysteinyl-tRNA synthetase | 229 |
| HP0888      | Iron(III) dicitrate ABC transporter, ATP-binding protein | 468 |
| HP0970      | Nickel-cobalt-cadmium resistance protein | 317 |
| HP0974      | Phosphoglycerate mutase | 257 |
| HP1059      | Holliday junction DNA helicase | 319 |
| HP1099      | 2-Keto-3-deoxy-6-phosphogluconate aldolase | 482 |
| HP1179      | Phosphopentomutase | 245 |
| HP1253      | Tryptophanyl-tRNA synthetase | 376 |
| HP1257      | Orotate phosphoribosyltransferase | 478 |
| HP1406      | Biotin synthetase | 361 |
| HP1480      | Seryl-tRNA synthetase (spot contains a mix of three proteins) | 263 |

* TIGR, The Institute for Genomic Research.

---

\[ \text{Proteomics of } H. \text{ pylori Cellular Proteins and Antigens} \]

\[ \text{Molecular \& Cellular Proteomics 2.12} \]

\[ \text{1281} \]
electrophoresis or in vivo, or represent just contaminations. It is important to notice that the GroEL peptide distribution was not fully reproduced within the three replicates. This might be caused by differences among the gel runs, or it might be a consequence of the low GroEL content within these spots so that in some cases these peptides might have fallen below the detection limit. Another possibility was that the criterion to find at least three out of the five most intense GroEL peaks was not rigid enough and that not all of these spots truly contain this protein. According to the identification results on average 1.6 spots/ORF were found in our dataset.

The protein alkyl hydroperoxide reductase (HP1563) was found in eight different spots. According to the position in the gel one spot had an apparent molecular weight that was double the weight of the main spot of the protein. Evidence was found that this spot contained dimers of the protein because cysteine-containing peptides were not found in the dimer spot (Fig. 5). This finding raised the question of whether these dimers exist in vivo or were artifacts of the two-dimensional gel electrophoresis. Artificial dimerization during the run of the second dimension can be ruled out because there was no smearing to be seen on the gels. As dimerization has little effect on the pI it could have taken place during the first run. The fact that no dimers were found from other members of the peroxiredoxin family form homodimers or because only a small part of the protein content of the gel one spot had an apparent molecular weight that was double the weight of the main spot of the protein. Evidence was found that this spot contained dimers of the protein because cysteine-containing peptides were not found in the dimer spot (Fig. 5). This finding raised the question of whether these dimers exist in vivo or were artifacts of the two-dimensional gel electrophoresis. Artificial dimerization during the run of the second dimension can be ruled out because there was no smearing to be seen on the gels. As dimerization has little effect on the pI it could have taken place during the first run. The fact that no dimers were found from other members of the peroxiredoxin family form homodimers or even decamers (24).

The protein γ-glutamyltranspeptidase (HP1118) was identified in two distinct spots, which were positioned far apart. The PMF-covered amino acid sequences of these spots were exclusive; their combined apparent masses added up to the theoretical mass calculated from the ORF so that we concluded that two fragments occurred (Fig. 6). Although both spots were only weakly antigenic in our immunoblots (11) the protein is known to be a virulence- and apoptosis-inducing factor of H. pylori that occurs in the form of two fragments (25, 26). Additionally, it was hypothesized that the protein is membrane-associated (25), and here the first 36 amino acids were not covered in the PMFs so that a cleavage of a signal sequence might have occurred.

A certain protein can not only be found in several spots, but a spot can also contain several proteins in the form of protein mixtures (similar amounts of protein), as minor components, or in the form of neighbor spot contaminants. In immunoproteomics antibody recognition of proteins separated on 2-DE gels is detected. Because highly specific antibodies may recognize very small amounts of protein it cannot be ruled out that minor components of spots might be detected instead of the major component. Therefore, one has to be sensitive to the identification of such antigens. Here we closely investigated 24 known antigenic spots as to whether they contained minor components using MS-Screener and hierarchical clustering (Table IV). Nine spots possibly contain other components, six were supposedly free of such components, and a further nine contained unknown peaks. From the nine spots first mentioned seven contained peptide masses from known antigens. However, only three spots showed concurrent recognition of the spot and the main spot of its minor component in our immunoblots (11) and might therefore have had an influence on the antigenicity. Two of these spots contain major components that were also found in other “clean” antigenic spots. Consequently, only one protein (spot 278, protease HP1012) remains that could have erroneously been assigned to be antigenic.

As mentioned above, the antigenic protein GroEL was identified in 15 different spots, and three of the five most intense peaks were found in a further nine spots. Twenty-one of these spots were recognized conjointly in the immunoblots (see Fig. 4 for the 15 GroEL-identified spots). For these spots no evidence for differential antigenicities of different protein species was found.

In our recent study (11) we identified five different groups of patients by hierarchical clusterings of immunoblot data. One criterion for the definition of two groups was the recognition of a spot cohort (spots 225, 226, 231, 232, 233, and 234), which was now identified to contain species of GroEL that are supposedly C-terminally truncated. For the reason that GroEL is a highly conserved antigen and that all of its known protein species were conjointly recognized by the sera of the patients, the biological relevance of these two patient groups remains unclear. Spot 154, which was a candidate for differential immunogenicity of different protein species of AtPA in the study mentioned above, was here identified to contain a mix of GroEL and AtPA. Several spots in this region (spots 154–157) lie side-by-side and contain either one of these proteins or mixtures of both so that in this case the identification that depends on spot assignment between immunoblots and 2-DE gels remains uncertain. A differentiation between GroEL and AtPA could be obtained by incubation of recombinant proteins with patient sera.

A dataset of 960 PMFs was used to compare automatic and manual data acquisition and investigate protein distributions in 2-DE gels. Large datasets can quickly be generated and identified with automatic procedures. For this purpose it is highly recommended to investigate replicate datasets to raise the rate of identification and improve reliability. Additionally, optimization of peak detection and database search parameters as well as calculation and removal of contaminants were shown to be advantageous. Manual measurements are still up-to-date especially for faint spots where procedures can be adapted individually. In addition we confirmed that immuno
Proteomics is a powerful hypothesis-free approach to find antigen candidates given that spot identification is performed cautiously.

Acknowledgments—We thank Ursula Zimny-Arndt for the preparation of excellent 2-DE gels, Klaus-Peter Pleissner for assistance with software, and Dietmar Waidelich for mass spectrometry support.

* This work was supported by Grants 031U/107 and 031U/207 from the Bundesministerium für Bildung und Forschung of Germany. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ To whom correspondence should be addressed: Max Planck Institute for Infection Biology, Campus Charité-Universitätsmedizin Berlin, Schumannstrasse 21/22, 10117 Berlin, Germany. Tel.: 49-30-28460507; E-mail: krah@mpiib-berlin.mpg.de.

REFERENCES

1. Jungblut, P. R., Bumann, D., Haas, G., Zimny-Arndt, U., Holland, P., Lamer, S., Siejak, F., Aebischer, A., and Meyer, T. F. (2000) Comparative proteome analysis of Helicobacter pylori. Mol. Microbiol. 36, 710–725

2. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzgerald, L. M., Lee, N., Adams, M. D., Venter, J. C., et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388, 539–547

3. Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uría-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vosis, G. F., and Trust, T. J. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 397, 176–180

4. Walker, M. M., and Crabtree, J. E. (1998) Helicobacter pylori infection and the pathogenesis of duodenal ulceration. Ann. N Y. Acad. Sci. 859, 96–111

5. Peek, R. M., Jr., and Blaser, M. J. (2002) Helicobacter pylori and gastrointestinal tract adenocarcinomas. Nat. Rev. Cancer 2, 28–37

6. Sabarth, N., Lamer, S., Zimny-Arndt, U., Jungblut, P. R., Meyer, T. F., and Bumann, D. (2002) Identification of surface proteins of Helicobacter pylori by selective biotinylation, affinity purification, and two-dimensional gel electrophoresis. J. Biol. Chem. 277, 27896–27902

7. Ull, M., Nilsson, I., Ljung, A., and Wadstrom, T. (2002) Identification of novel immunogenic proteins of Helicobacter pylori by proteome technology. J. Immunol. Methods 259, 1–10

8. Bumann, D., Aksu, S., Wendland, M., Janek, K., Zimny-Arndt, U., Sabarth, N., Meyer, T. F., and Jungblut, P. R. (2002) Proteome analysis of secreted proteins of the gastric pathogen Helicobacter pylori. Infect. Immun. 70, 3396–3403

9. Krah, A., and Jungblut, P. R. (2003) In Methods in Molecular Medicine: Molecular Diagnosis of Infectious Diseases (Decker, J., and Reischl, U., eds) Vol. 94, pp. 19–32, Humana Press Inc., Totowa, NJ

10. Haas, G., Kairali, G., Ebermayr, K., Metzger, W. G., Lamer, S., Zimny-Arndt, U., Dieser, S., Goebel, U. B., Vogt, K., Roznowski, A. B., Wiedemann, B. J., Meyer, T. F., Aebischer, T., and Jungblut, P. R. (2002) Immunoproteomics of Helicobacter pylori infection and relation to gastric disease. Proteomics 2, 313–324

11. Krah, A., Miehle, S., Pleissner, K. P., Zimny-Arndt, U., Kirsch, C., Lohn, N., Meyer, T. F., Jungblut, P. R., and Aebischer, T. (2004) Identification of candidate antigens for serologic detection of Helicobacter pylori infected patients with gastric carcinoma. Int. J. Cancer 105, 456–463

12. Kimmel, B., Bosserhoff, A., Frank, R., Gross, R., Goebel, W., and Beier, D. (2000) Identification of immunodominant antigens from Helicobacter pylori and evaluation of their reactivities with sera from patients with different gastroduodenal pathologies. Infect. Immun. 68, 915–920

13. McAtee, C. P., Lim, M. Y., Fung, K., Veligan, M., Fry, K., Chow, T., and Berg, D. E. (1998) Identification of potential diagnostic and vaccine candidates of Helicobacter pylori by two-dimensional gel electrophoresis, sequence analysis, and serum profiling. Clin. Diagn. Lab. Immunol. 5, 537–542

14. Jungblut, P., and Klose, J. (1985) Genetic variability of proteins from mitochondria and mitochondrial fractions of mouse organs. Biochem. Genet. 23, 227–245

15. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999

16. Schmidt, F., Schmid, M., Mattow, J., Faciuc, A., Pleissner, K.-P., and Jungblut, P. R. (2003) Iterative Data Analysis is the Key for Exhaustive Analysis of Peptide Mass Fingerprints from Proteins Separated by Two-Dimensional Electrophoresis. J. Am. Soc. Mass Spectrom. 14, 943–956

17. Klose, J., and Kobalz, U. (1995) Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. Electrophoresis 16, 1034–1059

18. Doherty, N. S., Littman, B. H., Reilly, K., Swindell, A. C., Buss, J. M., and Anderson, N. L. (1998) Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 353–363

19. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567

20. Bumann, D., Meyer, T. F., and Jungblut, P. R. (2001) Proteome analysis of the common human pathogen Helicobacter pylori. Proteomics 1, 473–479

21. Keller, B. O., and Li, L. (2000) Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures. J. Am. Soc. Mass Spectrom. 11, 88–93

22. Sabarth, N., Hurwitz, R., Meyer, T. F., and Bumann, D. (2002) Multiparameter Selection of Helicobacter pylori Antigens Identifies Two Novel Antigens with High Protective Efficacy. Infect. Immun. 70, 6499–6503

23. Ding, Q., Xiao, L., Xiong, S., Jia, Y., Que, H., Guo, Y., and Liu, S. (2003) Unmatched masses in peptide mass fingerprints caused by cross-contamination: An updated statistical result. Mol. Cell. Proteomics 2, 315–322

24. Wood, Z. A., Poole, L. B., Hantgan, R. R., and Karplus, P. A. (2002) Dimers in mass spectrometry. J. Am. Soc. Mass Spectrom. 13, 1163–1171

25. McCammon, J. A., Freire, C., and Casadevall, A. (1998) Gamma-Glutamyltransferase is a Helicobacter pylori virulence factor but is not essential for colonization. Infect. Immun. 66, 4168–4173

26. Harris, W. A., Janecki, D. J., and Reilly, J. P. (2002) Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 16, 1714–1722