Visualization and Analysis of Molecular Scanner Peptide Mass Spectra

Markus Müller, Robin Gras, and Ron D. Appel*
Swiss Institute of Bioinformatics, Geneva, Switzerland

Willy V. Bienvenut and Denis F. Hochstrasser*
Clinical Chemistry Laboratory, Geneva University Hospital, Geneva, Switzerland

The molecular scanner combines protein separation using gel electrophoresis with peptide mass fingerprinting (PMF) techniques to identify proteins in a highly automated manner. Proteins separated in a 2-dimensional polyacrylamide gel (2-D PAGE) are digested in parallel and transferred onto a membrane keeping their relative positions. The membrane is then sprayed with a matrix and inserted into a matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer, which measures a peptide mass fingerprint at each site on the scanned grid. First, visualization of PMF data allows surveying all fingerprints at once and provides very useful information on the presence of chemical noise. Chemical noise is shown to be a potential source for erroneous identifications and is therefore purged from the mass fingerprints. Then, the correlation between neighboring spectra is used to recalibrate the peptide masses. Finally, a method that clusters peptide masses according to the similarity of the spatial distributions of their signal intensities is presented. This method allows discarding many of the false positives that usually go along with PMF identifications and allows identifying many weakly expressed proteins present in the gel. (J Am Soc Mass Spectrom 2002, 13, 221–231) © 2002 American Society for Mass Spectrometry

At present, as complete genomes for an increasing number of organisms are available, attention must be focused on proteins encoded by the genes. In contrast to the static genome, the proteome of an organism is a highly dynamic and connected network, and new analytical methods have to be developed in order to describe its spatial and temporal changes and interactions [1]. An important step in this task is the high throughput identification of proteins, which nowadays mostly relies on efficient protein separation, mass spectrometry, protein sequence databases as well as bioinformatics [2].

One of the most important methods for protein separation is 2-dimensional polyacrylamide gel electrophoresis (2-D PAGE) [3]. This technique allows separating simultaneously thousands of proteins according to their isoelectric point (p*I*) and molecular weight (M_r) and displaying them on a 2-D map. Mass spectrometry (MS) has become one of the most powerful techniques to identify organic molecules. Among various applications, peptide mass fingerprinting (PMF) is frequently used because, combined with matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry [4, 5] it provides a rapid and sensitive method for protein identification. PMF compares the list of experimental masses of peptides, the peptide mass fingerprint, obtained by specific endoproteolytic digestion of proteins with the theoretical mass values calculated by in silico digestion of protein sequences. A valuation score shows how well the theoretical masses match the fingerprint [6–10]. Gras et al. [11] presented a PMF identification algorithm which is based on a scoring schema that takes into account important parameters like mass accuracy, protein coverage by matching peptides, number of missed cleavage sites, and the deviation of the measured p*I* and M_r values (if available) from theoretical predictions. In order to learn the weights of these parameters for the PMF identification score, a set of 91 PMF test spectra was used and optimal values of these weights were calculated by means of a genetic algorithm. Eriksson et al. [12] investigated the influence of different experimental parameters on statistical thresholds used to discern false matches for two different scoring schemas. Since the experimental mass fingerprint can match the theoretical peptide masses of a protein by chance, there is always a certain probability for false identifications in PMF. There is a trade-off between sensitivity and specificity of a database search: If the search is too restrictive, it might miss some proteins (false negatives), and if it is not restrictive

Published online January 16, 2002

Address reprint requests to Dr. M. Müller, CMU, Swiss Institute of Bioinformatics, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland. E-mail: markusmueller@isb-sib.ch

*Also at the University of Geneva, Geneva, Switzerland.

1044-0305/02/$20.00
PII S1044-0305(01)00358-0

Received August 29, 2001
Revised November 16, 2001
Accepted November 16, 2001

enough, it might find too many erroneous matches (false positives).

The precision of mass measurements certainly influences the sensitivity and specificity of PMF identification. Since the resolution of mass spectrometers has improved, calibration errors are now the limiting factor. These errors originate from uncertainties in the estimation of experimental parameters such as electric field strengths and initial ion velocities. Calibration of mass spectra is not a trivial problem even if internal standards are used. For TOF instruments, the function that relates the flight time with the *m/z* value and the algorithm to calculate the calibration parameters have to be carefully chosen in order to get a good precision. Christian et al. [13] described a method that is based on physical flight time equations [14, 15] and a simplex method to search for the optimal instrument parameters. This approach proved to be more robust than usual curve fitting methods, especially in the mass range where no standard masses were available.

Several partially automated methods have been proposed to excise protein spots from a stained gel, to submit the excised material to endoproteolytic digestion and to extract peptides from the excised gel [16]. The peptides are then loaded onto a MALDI sample plate and introduced into a mass spectrometer for PMF acquisition [17]. These methods have the inconvenience that the location of protein spots must be known prior to excision, and that the excision precision is limited (>1 mm). Recently, Binz et al. [18] introduced a new and highly automated approach, dubbed the molecular scanner, which combines 2-D PAGE separation techniques with PMF methods. In this approach, the proteins were digested firstly in the gel itself and then during transfer onto a collecting polyvinylidene fluoride (PVDF) membrane [19]. This membrane was sprayed with a matrix solution (α-Cyano-4-hydroxy cinnamic acid), and the co-crystallisation of the matrix and the peptides allowed MALDI-MS analysis. Since diffusion in this process was not relevant, the location of the peptides on the PVDF membrane corresponded to the location of their proteins in the gel [18]. The membrane was then scanned by a MALDI-TOF mass spectrometer. For each scanned point the acquired peptide mass fingerprint was submitted to a PMF identification program, which returned a list of matching proteins. A threshold that was based on a statistical analysis of erroneous identifications was used to distinguish false identifications by their average identification score [2]. This method provided good results for the most abundant proteins, but it had difficulties to distinguish weakly expressed proteins from noise. A graphical display allowed visualising the matching proteins on a 2-D map.

High throughput methods can produce a large amount of mass spectrometric data, and multidimensional visualization of these data is becoming more and more important. It allows surveying data and provides ideas for algorithmic solutions. One example is secondary ion mass spectrometry (SIMS) techniques, where natural tissues can be scanned with a spatial resolution of less than 100 nm and the resulting spectra can be used to visualize the 2- or 3-D distributions of secondary ions [20]. Stoeckli et al. [21] coated frozen thin sections of tissue with a solution of MALDI matrix, then dried and introduced them into a mass spectrometer, which scanned the sample. For a human brain tissue, an area of 8.5 mm $\times$ 8 mm was scanned with a grid spacing of 100 μm and the position of 45 ions were recorded and rendered as 2-D images.

In this paper, visualization of all mass fingerprints provides important information on the presence of chemical noise that is shown to be a potential source for false matches in the PMF identification procedure. The correlation of neighbouring spectra is used to recalibrate the mass fingerprints. In order to simplify PMF identifications, an algorithm calculates distributions of peptide signal intensities and joins the masses with similar distributions into clusters. These clusters represent protein spots, and many of them yield a clear PMF identification. These methods were developed in the framework of the molecular scanner, but we think that they are of more general interest since they deal with issues such as chemical noise, calibration, weak signal detection, and how contextual information can be used to improve results.

Methods

In this experiment, 1 mg *E. coli* proteins were separated by 2-D PAGE. After in-gel digestion, the proteins were submitted to a digestion-transfer and trapped on a PVDF membrane (Bio-Rad, Richmond, CA). A portion with a size of approximately 9 $\times$ 13 mm (corresponding to a p*I* range of 5.1–5.2 and a M_r range of 35'000–45'000 Da) was cut out from the membrane and pasted on the sampling plate of a MALDI-TOF mass spectrometer (Voyager Elite, Applied Biosystems, Framingham MA), which was equipped with a 337 nm UV laser. 5 mg/mL of α-cyano-4-hydroxycinnamic acid (4-HCCA from Sigma, St-Louis, MO) dissolved in 70% methanol was sprayed on the PVDF membrane. Then the membrane was scanned on a 48 $\times$ 32 grid with a sampling distance of 0.25 mm. 64 laser shots were fired at a frequency of 3 Hz leading to an acquisition time of about 9 h. The disc space needed to store all the spectra was 350 MB, which could be compressed to 3MB after peptide signal detection if just the mass fingerprints were stored. More details of the molecular scanner experiment discussed in this article can be found in [19].

The algorithms used for peptide signal detection and the PMF identification program SmartIdent are described in [11]. Since the concentration of some proteins was low, only a few of their peptide masses were detectable and the minimal number of matching masses for the PMF search was set to two if deconvoluted peptide mass lists were used and to three otherwise (since the standard version of SmartIdent requires at

least three matching masses, it was adapted to the needs of this experiment). The number of missed cleavages was set to one and only chemical modifications of cysteine and methionine were considered. The mass tolerance was set to 200 ppm. A reduced version of Swiss-Prot (Release 39.22 of 20 June 2001) that contained all 4740 proteins from *E. coli* was searched for PMF identification.

Calculations were performed on a 500 MHz Pentium processor with 128 MB RAM on Windows NT. Programs were written in C++ and Virtual Reality Modeling Language (VRML 2.0, http://www.sdsc.edu/vrml) was used for visualization. VRML is a software standard that defines the format of data files sent over the Internet for visualization and animation, and is therefore supported by Internet browsers. Netscape Communicator 4.7 was used to render the VRML data files and *m/z*-software by Proteometrics to render single spectra.

Results and Discussion

Visualization of Spectra

The data obtained in the molecular scanner experiment consisted of a set of mass spectra: One for each scan point. The first aim was to get an idea of how the data were structured. Since there were 1536 spectra, it was impossible to inspect and compare them by means of conventional visualization tools that are only able to render a few spectra at a time. We designed a method that allows circumventing this problem and inspecting all spectra at once.

Each mass detected in a spectrum can be associated with a point in a 3-D space (Figure 1a) where the horizontal plain corresponds to the scanned membrane and the vertical axis to the mass value. In Figure 1b all masses between 800 Da and 1000 Da are marked as points revealing that some masses were detected on a contiguous region of the scanned membrane, while others were found only on isolated lattice sites. For the main part of this paper we considered only masses that could be reproducibly detected in a neighborhood, because this provided more reliable results than working with all masses. Therefore a filter discarded a mass from a mass fingerprint if it could not be detected in the majority of the eight surrounding sites. All lattice sites were treated simultaneously and this process was repeated until a stable configuration was obtained, i.e. the filter can be represented as a synchronous cellular automaton [22]. This filter is different from a filter that selects the most intense peptide signals in an isolated spectrum since it takes into account the spatial correlation of the data. There were several low intensity peptide signals detected on a contiguous region that proved to be essential for the identification of a protein. The masses that pass this filter and do not belong to chemical noise (see below) are called contiguous masses and are depicted in Figure 1c.

Chemical Noise

Figure 1b reveals an interesting feature: Some masses cover the entire membrane while others are localized in spots. Figure 1d shows that the localized peptide signals at 951.5 Da and 999.7 Da are not distinguishable from ubiquitous masses at 804.4 Da, 820.4 Da, 838.2 Da and 936.1 Da by means of signal intensity. Figure 4 shows that signal intensity distributions of ubiquitous masses are flat in contrast to *E. coli* peptide masses. In order to automatically find ubiquitous masses, a routine tests how even and spread out an intensity distribution is. Therefore it divides the membrane into 26 regions of equal size (8×8) and calculates the deviation between each region's mean intensity $\bar{I}_i$ and the overall mean intensity $\bar{I}_{tot}$. If the sum over all regions of the relative deviations $\Sigma_i|\bar{I}_i - \bar{I}_{tot}|/\bar{I}_{tot}$ is smaller than a certain threshold (=20) and if the mass is detected on more than 72 sites, it is called ubiquitous (Table 1).

Since diffusion is limited in the molecular scanner technique [18], and since none of the ubiquitous masses (exception: 820.4 Da) could be associated with peptide masses of proteins annotated in the respective portion of the master SWISS-2DPAGE [23] gel (Swiss-Prot entries: IDH_ECOLI, METK_ECOLI, PGK_ECOLI, ACEA_ECOLI), these ubiquitous masses do not stem from proteins of the *E. coli* sample. However, some of these ubiquitous masses could be attributed to known impurities from tryptic autolysis and various forms of human keratin, whereas the remaining masses could stem from modified or unknown impurity peptides and matrix clusters. Matrix clusters form another source of chemical noise in the low mass range, especially if the amount of protein to be analyzed is low [24, 25], but in contrast to contaminating peptides their mass and intensity are not reproducible and it is not sure whether they could be detected over the entire membrane. In addition, the ubiquitous masses could not be explained by a formula for matrix cluster masses as described in [24]. Whatever the source for the masses listed in Table 1 is, it would be impossible to discern them from low intensity peptides from the *E. coli* sample without the knowledge of their spatial distribution provided by the molecular scanner data.

Calibration

Masses detected over the entire membrane could be used to investigate the calibration of the mass spectrometer. Figure 2a reveals that mass values were locally quite stable, but varied significantly over the entire membrane, whereas the difference between the minimal and maximal measured value of the trypsin peptide mass at 842.509 Da was about 1 Da because the membrane was warped at its upper edge (high M_r values), and because physical conditions as electric field strength depend on the position of the sampling plate [26]. Therefore it was impossible to assign precise mass values useful for all spectra, and a large mass deviation

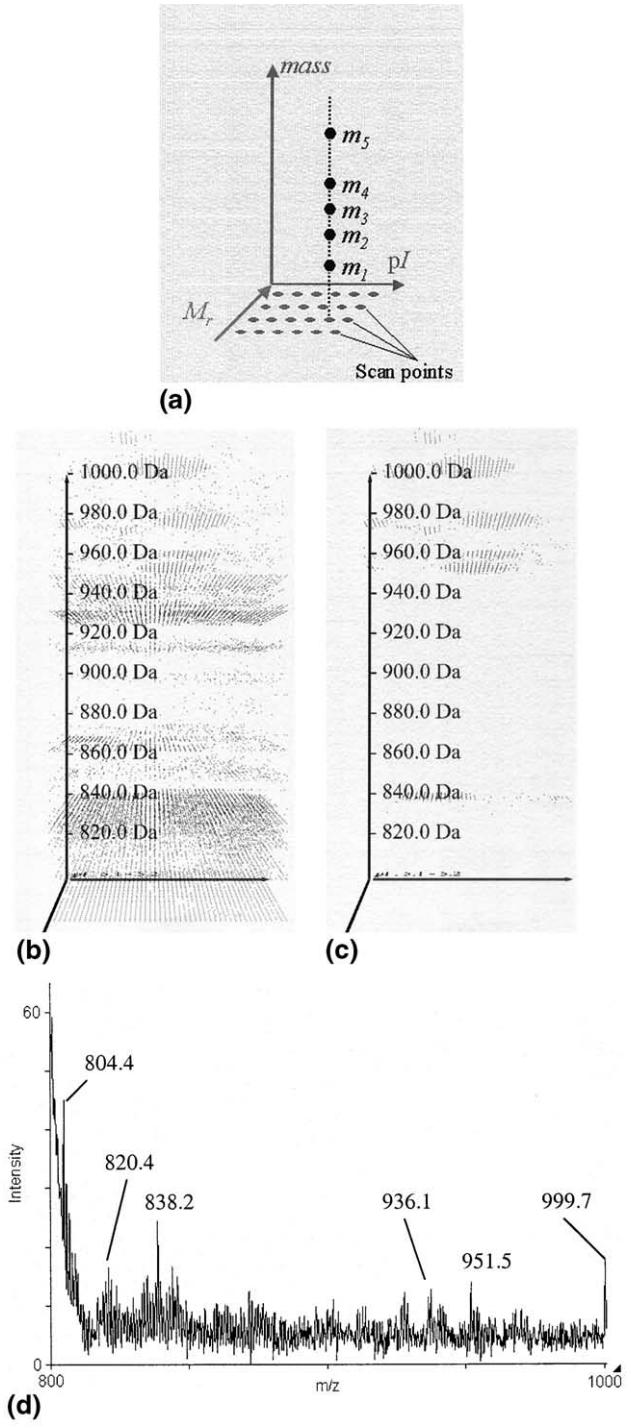

Figure 1. (**a**) The pI axis goes from 5.1 to 5.2, whereas the M_r axis is inverted and goes from 45'000 Da to 35'000 Da. Masses of one spectrum (m_1,. . .,m_5) are schematically depicted. (**b**) Masses between 800 and 1000 Da. The peptide signal detection threshold was set to the optimal value used for the identification where also small signals are detected (signal height >2.2*noise). (**c**) Contiguous masses between 800 and 1000 Da. Only the masses that were detected in a contiguous, but well localized region are shown. (**d**) 800 Da –1000 Da portion of a spectrum from the upper right part of the scanned membrane. Only an arbitrary selection of detected peptide signals is labeled.

Table 1. Ubiquitous masses

Mass (Da)[a]	Number of sites[b]	Alleged origin[c]
804.5	1320	Keratin[2]
820.4	761	
823.4	112	Keratin[1,2]
829.3	139	
832.5	377	Keratin[2]
833.4	265	
834.4	451	
838.3	636	
839.3	315	
842.5	1251	Trypsin
845.3	665	
859.5	202	
861.3	97	Keratin[1]
871.2	577	Keratin[1]
912.4	234	
913.5	103	
914.5	74	
926.4	868	
927.5	188	
936.2	355	
940.5	582	
1027.2	305	
1032.6	154	
1045.7	366	Trypsin
1046.6	180	Keratin[2]
1060.3	105	Keratin[1]
1092.2	234	
1126.7	170	Keratin[1]
1164.7	206	
1480.0	72	
1804.1	99	
1994.3	93	Keratin[2]
2118.4	136	Keratin[1]
2211.4	92	Trypsin
2250.2	89	

[a]Ubiquitous mass value. Since this value is not exactly the same in all spectra where the mass was found the median value is displayed (after calibration, see below).
[b]The number of spectra where the mass was found (maximal 1536). Sometimes, masses were detected with a deviation of about +1 Da from a keratin/trypsin peptide mass. This might be due to difficulties to detect the monoisotopic mass for very small peptide signals.
[c]If a mass matched a trypsin (SwissProt entry: TRYP_PIG) or a keratin peptide, it is indicated in this field (one missed cleavage, maximal mass deviation 200 ppm). The following human keratins produced more than one match: 1) K1CM_HUMAN, 2) K2C1_HUMAN.

of 700 ppm about the median values had to be taken into account. A re-calibration of the spectra would facilitate data handling, and we had to device a method that does not rely on internal standard masses since these were not used in the experiment described here.

Since we had no information about flight times and how they had been converted into mass values, it was not possible to apply the method described in [13] to our problem and we had to guess a function that calculates the corrected masses from the original masses. Egelhofer et al. [26] used a linear relationship, which was a reasonably good approximation to their data and is easy to calculate with. We chose a different approach:

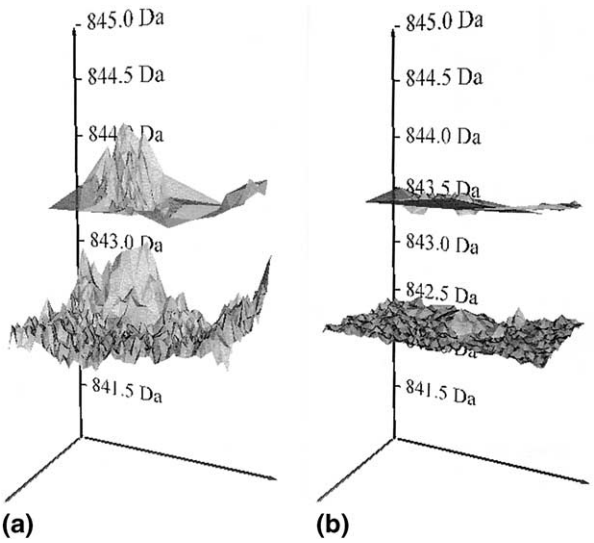

Figure 2. (**a**) Masses between 841 and 845 Da. The masses around 842.5 Da, which are detected over the entire membrane, correspond to a trypsin peptide, whereas the masses around 843.5 Da stem from isocitrate lyase (Swiss-Prot entry ACEA_ECOLI) and are localized in the pI-M_r plane except for a few outliners. The scattering of mass values is due to calibration errors that become larger (0.7 Da) towards the edges of the membrane. For better visualization, the mass values are rendered as a surface plot. (**b**) As in (**a**), but after calibration using the algorithm described in the text.

$$m^{1/2}{}_{corrected} = a_1 m^{1/2} + a_2 m + a_3 m^{3/2} + a_4 m^2 + k \qquad (1)$$

which took into account additional terms which fit well to observed data (not shown). If the calibration correction is known for a set of masses $\{m^i, m^i{}_{corrected}\}$, then the parameters a_k can be calculated by a robust fit [27] of eq 1.

None of the trypsin or keratin peptide masses could be detected in all spectra and therefore they could not be used as internal standards. However, many masses found in one spectrum could also be detected in the spectra of the neighboring scan points with a relatively small mass deviation (<200 ppm), which would allow at least a relative adjustment of masses of a spectrum with respect to its neighbors. If there was a way to calibrate some master spectra, the relative adjustment could be used to calibrate the remaining spectra.

Some scanned points provided very clear PMF identifications even if large mass deviations were allowed. The peptide masses of the identified proteins can then be used as standard masses for the calibration of the associated mass fingerprints. An iterative algorithm was then used to calibrate the remaining the spectra:

1. Choose some sites with very clear identifications and use the theoretical masses of the matching peptides as mass standards in order to calibrate the respective fingerprints with eq 1.
2. For each spectrum that was not calibrated in Step 1, one of the following steps is performed: (a) If a spectrum is found in a 3 × 3 neighborhood that has already been calibrated in Step 1, adjust the masses with respect to this spectrum, i.e., find the masses that are common in both spectra (with a mass tolerance of 200 ppm) and fit eq 1 to these values. If several such spectra are found, take an average adjustment, i.e., take the mean values of the parameters a_k. (b) If no such spectra are found, take the average adjustment with respect to all spectra in the 3 × 3 neighborhood. This step (Step 2) is performed simultaneously for all spectra and a new, corrected set of fingerprints is obtained that replaces the old fingerprints.
3. Repeat Step 2 until the variations of the masses over the membrane are small enough.

The result of this procedure is depicted in Figure 2b. 109 master spectra were selected, all in the upper part of the membrane where the abundant proteins were found. The remaining variation of the mass values over the entire membrane was smaller than 200 ppm. This method has one drawback: If no clear identifications could be found in an experiment, the calibration provides only a relative adjustment between neighbors. In this case the known masses of trypsin and keratin might serve as standards and nevertheless allow a recalibration of the mass fingerprints.

Identification and Clustering of Masses

A peptide mass fingerprint is usually contaminated with chemical noise and masses of fragmented or modified peptides. In addition, the resolution of the 2-D PAGE was limited and some spots overlapped and abundant proteins covered weakly expressed ones. On the sites of some weakly expressed proteins many of the detected peptide masses stemmed from their abundant neighbors, and the PMF identification of the spectrum obtained at the respective sites yielded a list of matches where the weakly expressed proteins were only found in a lower rank. All these fake masses strongly enhance the number of mass combinations and produce false matches in the database search.

Figure 3 shows SmartIdent scores if untreated fingerprints are submitted. Only the abundant proteins IDH_ECOLI, 6PGD_ECOLI, METK_ECOLI, and PGK_ECOLI were coherently detected with the highest score, other proteins scored highest only at isolated sites and disappeared elsewhere in the mist of false identifications. The protein ATOC_ECOLI has peptides at 804.428 Da, 820.423 Da, 842.404 Da, 1045.484 Da and 1046.603 Da, which match signals produced by chemical noise, and is therefore detected over a large part of the membrane (Figure 3b). Its p*I* of 6.01 and molecular weight of 52176.39 Da are outside the scanned portion of the membrane and it is unlikely to be found over such a large region. Though its score is significant it is a false identification. ACEA_ECOLI is a protein annotated in SWISS-2-D PAGE and it is identified over a contiguous region, but not with a very significant score (Figure 3c). Also YAGE_ECOLI is identified in the same

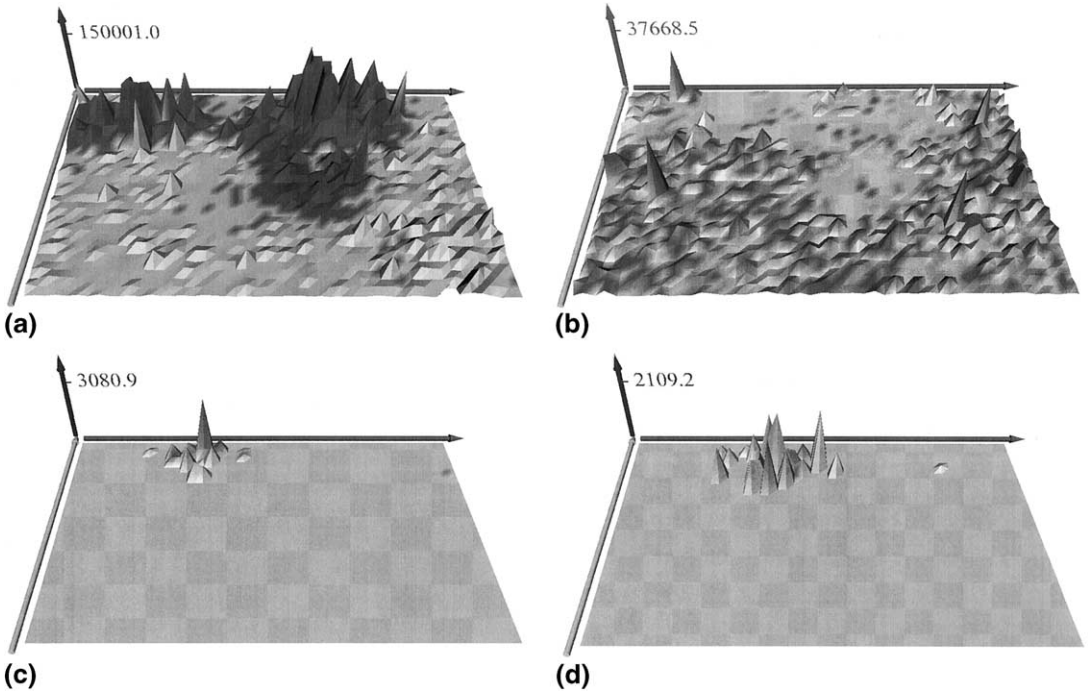

Figure 3. SmartIdent scores for calibrated but otherwise untreated spectra. (**a**) Highest score for each scanned site. For a better visualization, score values were cut at 150'000. Sites are dark if one of the proteins of Figure 5 was detected with the highest score and light otherwise. (**b**) Score of acetoacetate metabolism regulatory protein atoC (ATOC_ECOLI), which produced matches with chemical noise. Chemical noise is sometimes suppressed in spots of abundant proteins, which explains the holes in the score landscape. (**c**) Score of ACEA_ECOLI. (**d**) Score of hypothetical protein yagE (YAGE_ECOLI).

region with a similar score (Figure 3d) and it is difficult to decide whether it is a true or a false identification (a more detailed analysis as described below discards it as a false one). This shows that the identification score does not provide sufficient information for weakly expressed proteins and we have to investigate PMF identifications in further detail.

A peptide mass fingerprint in the overlapping zone of the spots of *isocitrate dehydrogenase* (IDH_ECOLI) and *s-adenosylmethionine synthetase* (METK_ECOLI) was sent to SmartIdent. The protein with the highest score was IDH_ECOLI (score: 818034.11; 13 matching peptides) followed by *allantoate amidohydrolase* (ALLC_ECOLI; score: 51341.17; 6 matching peptides) and others with slowly decreasing scores. METK_ECOLI (score: 25122.18; 5 matching peptides) was only found in the sixth rank. While the score of the first protein is significantly higher than the score of the second protein, there is almost no difference between the second and third rank, and it is very difficult to decide whether ALLC_ECOLI is an erroneous match without additional information. However, the signal intensities of the matching masses revealed interesting properties (Figure 4).

For METK_ECOLI all peptides except the one at 1155.684 Da showed a similar intensity distribution. The peptide at 1155.684 Da, which stemmed from the neighboring protein *phosphoglycerate kinase* (PGK_ECOLI), showed lower molecular weight and higher p*I* values in good correspondence to other peptides of PGK_ECOLI. The case of ALLC_ECOLI was very different because no intensity distribution specific to this protein could be found. The first two masses (804.419 Da and 820.413 Da) were not localized and were part of the chemical noise (Table 1). The next mass (951.456 Da) belonged to METK_ECOLI, and the peptide at 1193.623 Da was similar to the one at 1155.684 Da and also belonged to PGK_ECOLI, whereas the remaining two masses (1177.629 Da and 2086.002 Da) could be attributed to IDH_ECOLI. Therefore we assume that the identification of ALLC_ECOLI was erroneous.

This analysis identified two possible causes for erroneous identifications: Chemical noise and overlapping protein spots. Chemical noise could be identified using the method described above and purged from the mass fingerprints. In order to separate masses from overlapping proteins, the masses that had similar intensity distributions had to be identified and put into the same cluster. If each protein corresponds to a particular pattern of intensity distribution, then the clusters will only contain masses that stem from the same protein.

It is known that peptide signal intensity in MALDI-MS has a poor shot-to-shot reproducibility due to matrix/analyte inhomogeneity, variation in laser power, and detector nonlinearity [28]. Normalization of the signal intensity with internal standards improves reproducibility and allows a quantitative analysis over

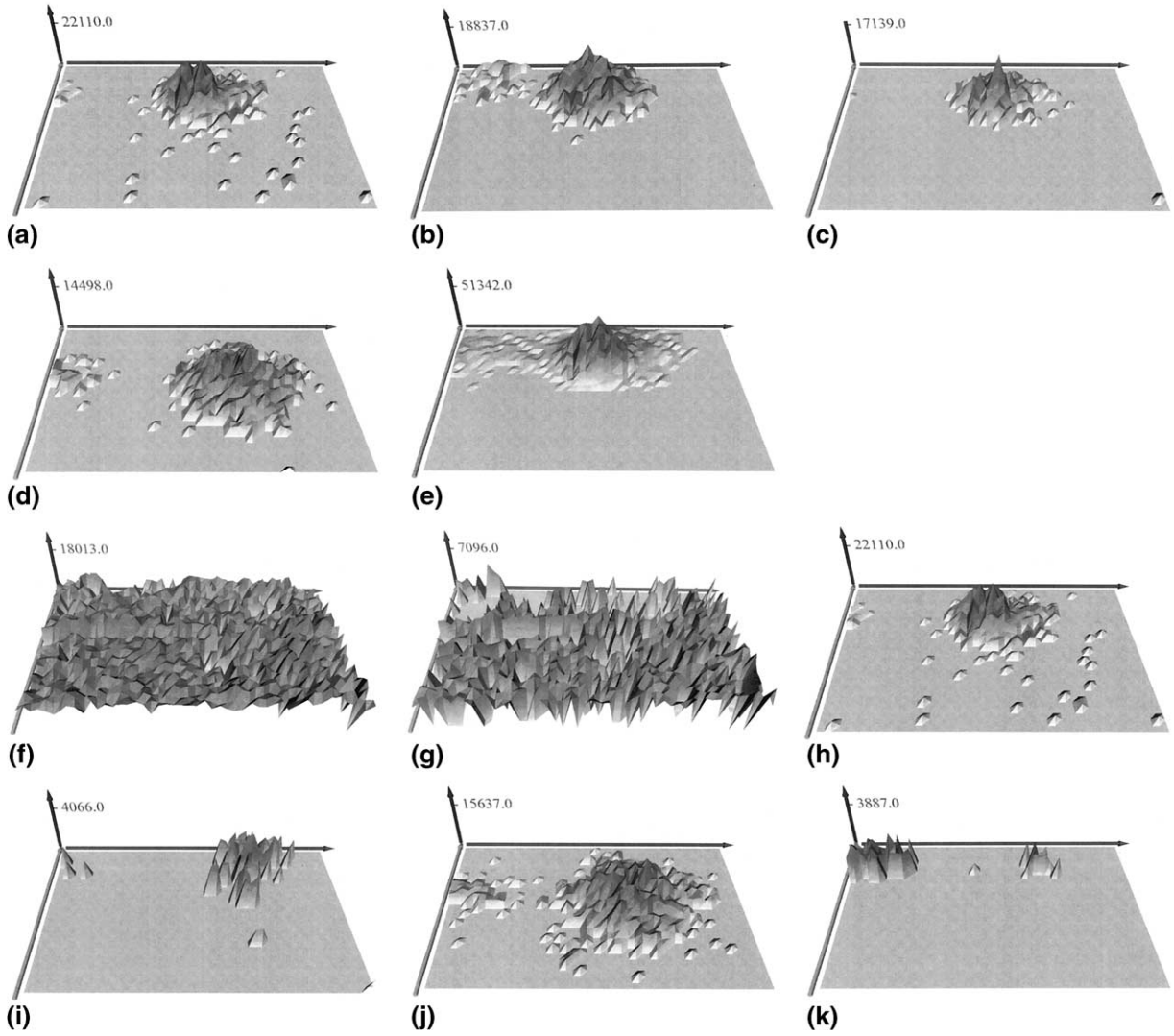

Figure 4. The vertical axis represents the peptide signal intensities (peptide signal heights) as a function of the position on the membrane. The intensity was set to 0 if no peptide mass could be detected at the respective position within ±100 ppm of the theoretical peptide mass. Note that the scale varies from case to case. (**a**)–(**e**) Intensity distribution of the matching peptides of METK_ECOLI. (**f**)–(**k**) Intensity distribution of the matching peptides of ALLC_ECOLI.

two orders of magnitude [29, 30], at least if the internal standard is chemically similar to the measured peptides and if concentrations are low enough to avoid suppression effects [31]. Kratzer et al. [32] investigated suppression effects in MALDI-MS with 4-HCCA as matrix, and obtained a good reproducibility of the absolute signal intensities after averaging over 50 laser shots. They showed with a mixture of 10 peptides that up to 2.5 pmol/peptide the absolute signal intensities of all peptides increased nearly linearly with increasing peptide amount, but for higher amounts, complicated nonlinear suppression effects came into play depending on the presence of basic amino acids, hydrophobicity, and peptide length. The longer, more hydrophobic and arginine containing peptides did not decrease in signal intensity in the measured range (100 fmol–25 pmol), but stayed constant or slightly increased for high concentrations, whereas other peptides were strongly suppressed. In the experiment discussed here and in similar experiments [18], the absolute signal intensity of the contiguously detected peptides always increased towards the center of a spot where peptide concentration is highest. For the amount of *E. coli* sample analyzed, the amount of protein in a spot is expected to be in the low pmol range producing digested peptides of even lower amount, which might well be in the linear range. For the well-expressed, contiguously detected peptides one can therefore assume that the signal intensity is positively correlated with the concentration of its protein in the gel, and a similar intensity distribution of two peptides indicates that they stem from the same protein.

In order to quantify the similarity between intensity distributions, a correlation measure had to be defined.

Table 2. Correlation between intensity distributions

	999.558 Da	1254.742 Da	1155.684 Da	1193.623 Da	2086.002 Da
999.558 Da	1.000	0.606	−0.084	−0.078	−0.097
1254.742 Da	0.606	1.000	−0.046	−0.007	−0.074
1155.684 Da	−0.084[a]	−0.046	1.000	0.747	−0.058
1193.623 Da	−0.078	−0.007	0.747	1.000	−0.036
2086.002 Da	−0.097	−0.074	−0.058	−0.036	1.000

[a]Negative values were close to 0 because peptides with a bad linear correlation usually had little overlap. A cut-off of 0.35 was used in the clustering algorithm.

We chose a modified version of the linear correlation that also takes into account how strongly two distributions overlap. The correlation between two masses m_i and m_j is defined by

$$corr_{ij} = \frac{2n_{ij}}{n_i + n_j} \frac{\frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (h_{ik} - \bar{h}_i)(h_{jk} - \bar{h}_j)}{\sigma_i \sigma_j}; -1 \le corr_{ij} \le 1 \quad (2)$$

$$\bar{h}_i = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} h_{ik};\ \sigma_i^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (h_{ik} - \bar{h}_i)^2;$$

where n_i (n_j) is the number of sites where m_i (m_j) is detected, respectively, and n_{ij} is the number of sites where both m_i and m_j are found. h_{ik} is the intensity of m_i at site k. The factor $2n_{ij}/(n_i + n_j)$ is 1 if m_i and m_j are found on exactly the same sites and 0 if there is no overlap at all. The sums in the above equations always go over all sites where both masses are detected. This correlation measure does not change if a signal is multiplied by a constant factor and it is stable against small local variations in absolute signal intensity as long as h_{ik}/h_{jk} are kept constant.

Table 2 shows the correlation value for some masses from Figure 4. Obviously there is a strong correlation between masses of the same protein and a negative correlation between masses belonging to different proteins.

We calculated the correlation (eq 2) between the 124 contiguous masses and performed a hierarchical cluster analysis [33] in order to group the masses according to their intensity distribution, which yielded 20 clusters, 11 of which contained more than two masses. Since the intensity of a mass should be highest where the concentration of protein is maximal, the summit of an intensity distribution should indicate the center of a spot. Figure 5 shows the summits of all 124 masses colored according to the cluster they belong to. It shows that the summits stemming from the same cluster lie close together unless the protein corresponding to the cluster was found on different spots. There was no overlapping of the centers of different clusters, and several weakly expressed spots could be well separated from their intense neighbors.

The algorithm described above provides a means of clustering the contiguous peptide masses that belong to the same spot, and the masses of the same cluster can be submitted to the PMF identification program. Since chemical noise was removed and all the masses stemmed from the same protein (assuming that the spot centers of neighboring proteins are sufficiently separated), these identifications should contain less erroneous matches. Instead of the 1536 mass lists of all scanned points, only the 20 mass lists of all clusters had to be submitted to SmartIdent.

Some of these mass lists had only a few entries and the identification score was not discriminative enough to clearly identify a protein. In this case the contiguous masses were not sufficient and we had to revert to the entire set of masses. Therefore all masses in a 3 × 3 neighborhood of the cluster center that appeared more than once were collected, and all the proteins that matched at least two contiguous masses were compared with these extended mass lists. If the extended mass list clearly distinguished a protein, this identification was accepted.

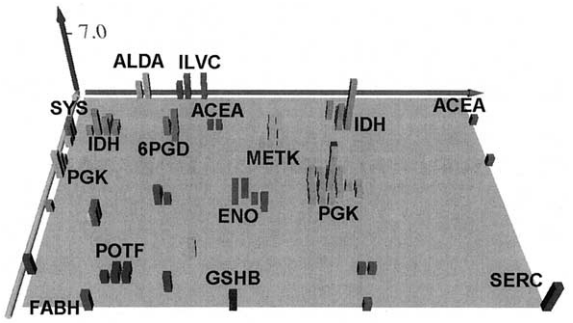

Figure 5. Summits of the intensity distributions of all 124 masses found on a contiguous but localized region. The intensity distributions were smoothed using a median filter before the summits were calculated. The vertical axis indicates the number of summits found on the respective scan point. The groups that could be identified (13 of 20) carry a label: Aldehyde dehydrogenase A (ALDA_ECOLI), ketol-acid reductoisomerase (ILVC_ECOLI), seryl-tRNA synthetase (SYS_ECOLI), isocitrate dehydrogenase (IDH_ECOLI), 6-phosphogluconate dehydrogenase (6PGD_ECOLI), isocitrate lyase (ACEA_ECOLI), s-adenosylmethionine synthetase (METK_ECOLI), phosphoglycerate kinase (PGK_ECOLI), enolase (ENO_ECOLI), putrescine-binding periplasmic protein [precursor] (POTF_ECOLI), 3-oxoacyl-[acyl-carrier-protein] synthase III (FABH_ECOLI), glutathione synthetase (GSHB_ECOLI), phosphoserine aminotransferase (SERC_ECOLI). IDH_ECOLI, PGK_ECOLI and ACEA_ECOLI were found on two spots.

Table 3. Identification results

Swiss-Prot entry[a]	Score[b]	Rank[c]	Best number of matched masses[d]	Number of contiguous masses[e]	Number of matched masses[f]
ALDA_ECOLI	1066.50 (59.64)	1	5	3	3 (3)
ILVC_ECOLI*	152.48 (274.84)	2	9	8	3 (15)
SYS_ECOLI	179.83 (26.83)	1	4	2	2 (3)
IDH_ECOLI	870993.90 (946.29)	1	15	17	11 (16)
6PGD_ECOLI	56913.04 (855.43)	1	9	7	5 (8)
ACEA_ECOLI*	116.18 (53.67)	1	4	2	2 (3)
METK_ECOLI	313362.28 (379.00)	1	9	12	7 (9)
PGK_ECOLI	117051.12 (4221.04)	1	11	30	8 (11)
ENO_ECOLI*	6889.07 (1941.52)	1	7	9	4 (8)
POTF_ECOLI	977.17 (0.02)	1	5	5	3 (3)
FABH_ECOLI	84.14 (4.59)	1	4	2	2 (3)
GSHB_ECOLI*	98.68 (172.43)	2	4	2	2 (5)
SERC_ECOLI	74.93 (2.15)	1	4	4	2 (3)

[a]Swiss-Prot entry: Swiss-Prot entry for the proteins that could be identified. Asterisks mark identifications that had to be verified by the extended mass lists as described in the text.
[b]SmartIdent identification score. If the protein was found in the first rank the value in parenthesis represents the score of the second rank, on the other hand, if the protein was not found in the first rank the value in parenthesis represents the score of the first rank.
[c]The highest number of matched masses of the respective protein found among the original 1536 mass fingerprints.
[d]Rank of the protein in the list of matching proteins sorted with respect to the score.
[e]Number of contiguous masses in a cluster that were submitted to SmartIdent.
[f]Number of contiguous masses of a cluster that matched peptides in the database search. The number of matching masses of the extended mass list (see text) is indicated in parenthesis.

Table 3 shows PMF identifications for those clusters that could be clearly identified. IDH_ECOLI, 6PGD_ECOLI, METK_ECOLI, and PGK_ECOLI, the most abundant proteins, had a high score, which was much higher than the score of the protein in the second rank. It is remarkable that many of the contiguous masses attributed to the clusters of these proteins did not match a peptide mass, the most extreme case being PGK_ECOLI with 22 unmatched masses. A visual examination revealed that these masses really had a similar intensity distribution, and the problem did not lie in the clustering algorithm. We see two different explanations: First, the peptides could be highly modified, fragmented or produced by unspecific cleavage and, second, other proteins could be present in the same spot. In the case of PGK_ECOLI the intensity distributions almost always showed two spots, and it seems unlikely that another protein could be present in the same two spots. In addition, there was no other protein that matched a lot of masses from the extended mass list. Therefore the first hypothesis seems more likely. Of the 22 unmatched masses only three could be explained by modified peptides (one carboxyamidomethyl cysteine, one dimethylation, and one phosphorylation), therefore unspecific cleavage or fragmentation seems to have caused most of these masses, but further investigation has to be carried out in order to give a definite answer.

Other identifications were less clear, but could be confirmed with the extended mass lists. Even if the protein database was small since it just contained the *E. coli* proteins, it is remarkable that some proteins could be identified with only two masses. Therefore, if the right masses are selected, a small number of masses might be sufficient for a clear PMF identification and we think that the algorithm presented here provides such a good selection. The groups that could not be identified still provided valuable information on the presence of a spot, which may be useful for gel matching.

Conclusion

The molecular scanner is a protein identification technique that is able to scan a gel without previous knowledge of spot locations. Since the distance between two points at which the membrane is sampled is smaller than the average spot size, several spectra per spot are obtained. This allows applying optimization methods that make use of the spatial correlation present in the data.

Visualization of all peptide mass fingerprint data revealed that some masses are localized in spots whereas other masses, especially in the lower mass region, spread out over the entire membrane. These masses were attributed to chemical noise and were discarded from the mass fingerprints. If only isolated spectra were available, the identification of chemical noise masses would be very difficult and these masses could disturb the PMF identification. Since the membrane was slightly warped after it had been pasted on the sampling plate of the spectrometer, and since the physical parameters that define the m/z value of peptides as a function of their flight time depended of the position of the sampling plate, the overall calibration of the spectra was bad. A few master spectra that permitted very clear PMF identifications could be calibrated using matched peptide masses as internal standards. The calibration of the remaining spectra was strongly

improved by using the correlation between neighboring spectra.

By selecting the masses that were detected on a contiguous, but limited region of the membrane, the noise in the data was reduced. The distributions of the peptide signal intensities of these masses seemed to reflect the concentration of the proteins they stemmed from. Masses with similar peptide signal intensity distributions were put together in clusters, which allowed separating masses that stemmed from overlapping proteins. 20 different clusters were obtained in this way and were submitted to the PMF identification program, which provided clear identifications for 13 of them.

These are only some applications that are possible with molecular scanner data. We are currently working on a new PMF identification scoring method that automatically takes into account the 2-D aspect of the data. A very intriguing prospect for future development comes from a new generation of mass spectrometers such as MALDI-TOF/TOF [34] and MALDI-QqTOF [35] machines, which could combine the MALDI scanning technique with MS/MS identification. The mass grouping method could then be used, after a first MS scan, to efficiently select parent masses for subsequent fragmentation analysis. A new technique [36], where the peptides are put on a porous silicon surface allowing disorption-ionization (DIOS) without a matrix directly from the surface, could also have a direct application in the framework of the molecular scanner.

Acknowledgments

This work was supported by the Swiss National Fund for Scientific Research (grant 31-52974.97) and the Helmut Horten Foundation. The authors would like to thank Pierre-Alain Binz, Salvo Peasano, and Jean-Charles Sanchez for their very useful contributions.

References

1. Godovac-Zimmermann, J.; Brown, L. R. Perspectives for Mass Spectrometry and Functional Proteomics . *Mass Spectrom. Rev.* **2001,** *20(1),* 1–57.
2. Bienvenut, W. V.; Müller, M.; Palagi, P.; Gasteiger, E.; Heller, M.; Jung, E.; Giron, M.; Gras, R.; Gay, S.; Binz, P. A.; Hughes, G. J.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. Proteomics and Mass Spectrometry: Some Aspects and Developments. *In Mass Spectrometry and Genomic Analysis.* Kluwer: The Netherlands, 2001; 1st ed.; pp 1–53.
3. Bjellqvist, B.; Ek, K.; Righetti, P. G.; Gianazza, E.; Gorg, A.; Westermeier, R.; Postel, W. Isoelectric Focusing in Immobilized pH Gradients: Principle, Methodology, and Some Applications. *J. Biochem. Biophys. Methods* **1982,** *6(4),* 317–339.
4. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins With Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60(20),* 2299–2301.
5. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. Protein and Polymer Analysis up to *m/z* 100,000 by Laser Ionization Time-of-Flight Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1988,** *2,* 151–153.
6. 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(11),* 5011–5015.
7. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein Identification by Mass Profile Fingerprinting. *Biochem. Biophys. Res. Commun.* **1993,** *195(1),* 58–64.
8. Mann, M.; Hojrup, P.; Roepstorff, P. Use of Mass Spectrometric Molecular Weight Information to Identify Proteins in Sequence Databases. *Biol. Mass Spectrom.* **1993,** *22(6),* 338–345.
9. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Curr. Biol.* **1993,** *3,* 327–345.
10. Yates, J. R. I.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide Mass Maps: A Highly Informative Approach to Protein Identification. *Anal. Biochem.* **2001,** *214,* 397–408.
11. Gras, R.; Müller, M.; Gasteiger, E.; Gay, S.; Binz, P. A.; Bienvenut, W.; Hoogland, C.; Sanchez, J. C.; Bairoch, A.; Hochstrasser, D. F.; Appel, R. D. Improving Protein Identification from Peptide Mass Fingerprinting Through a Parameterized Multi-Level Scoring Algorithm and an Optimized Peak Detection. *Electrophoresis* **1999,** *20(18),* 3535–3550.
12. Eriksson, J.; Chait, B. T.; Fenyo, D. A Statistical Basis for Testing the Significance of Mass Spectrometric Protein Identification Results. *Anal. Chem.* **2000,** *72(5),* 999–1005.
13. Christian, N. P.; Arnold, R. J.; Reilly, J. P. Improved Calibration of Time-of-Flight Mass Spectra by Simplex Optimization of Electrostatic Ion Calculations. *Anal. Chem.* **2000,** *72(14),* 3327–3337.
14. Juhasz, P.; Vestal, M. L.; Martin, S. A. On the Initial Velocity of Ions Generated by Matrix-Assisted Laser Desorption Ionization and Its Effect on the Calibration of Delayed Extraction Time-of-Flight Mass Spectra. *J. Am. Soc. Mass Spectrom.* **1997,** *8,* 209–217.
15. Vestal, M. L.; Juhasz, P. Resolution and Mass Accuracy in Matrix-Assisted Laser Desorption Ionization Time-of-Flight. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 892–911.
16. Lopez, M. F. Better Approaches to Finding the Needle in a Haystack: Optimizing Proteome Analysis Through Automation. *Electrophoresis* **2000,** *21(6),* 1082–1093.
17. Traini, M.; Gooley, A. A.; Ou, K.; Wilkins, M. R.; Tonella, L.; Sanchez, J. C.; Hochstrasser, D. F.; Williams, K. L. Towards an Automated Approach for Protein Identification in Proteome Projects. *Electrophoresis* **1998,** *19(11),* 1941–1949.
18. Binz, P. A.; Müller, M.; Walther, D.; Bienvenut, W. V.; Gras, R.; Hoogland, C.; Bouchet, G.; Gasteiger, E.; Fabbretti, R.; Gay, S.; Palagi, P.; Wilkins, M. R.; Rouge, V.; Tonella, L.; Paesano, S.; Rossellat, G.; Karmime, A.; Bairoch, A.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. F. A Molecular Scanner to Automate Proteomic Research and to Display Proteome Images. *Anal. Chem.* **1999,** *71(21),* 4981–4988.
19. Bienvenut, W. V.; Sanchez, J. C.; Karmime, A.; Rouge, V.; Rose, K.; Binz, P. A.; Hochstrasser, D. F. Toward a Cinical Molecular Scanner for Proteome Research: Parallel Protein Chemical Processing Before and During Western Blot. *Anal. Chem.* **1999,** *71(21),* 4800–4807.
20. Pacholski, M. L.; Winograd, N. Imaging with Mass Spectrometry. *Chem. Rev.* **1999,** *99,* 2977–3005.
21. Stoeckli, M.; Chaurand, P.; Hallahan, D. E.; Caprioli, R. M. Imaging Mass Spectrometry: A New Technology for the Analysis of Protein Expression in Mammalian Tissues. *Nat. Med.* **2001,** *7(4),* 493–496.
22. Toffoli, T.; Margolus, N. *Cellular Automata Machines.* MIT Press: Cambridge, 1987.
23. Hoogland, C.; Sanchez, J. C.; Tonella, L.; Binz, P. A.; Bairoch, A.; Hochstrasser, D. F.; Appel, R. D. The 1999 SWISS-2-D PAGE Database Update. *Nucleic Acids Res.* **2000,** *28(1),* 286–288.

24. Keller, B. O.; Li, L. Discerning Matrix-Cluster Peaks in Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectra of Dilute Peptide Mixtures. *J. Am. Soc. Mass Spectrom.* **2000,** *11(1),* 88–93.
25. Land, C. M.; Kinsel, G. R. The Mechanism of Matrix to Analyte Proton Transfer in Clusters of 2,5-Dihydroxybenzoic Acid and the Tripeptide VPL. *J. Am. Soc. Mass Spectrom.* **2001,** *12(6),* 726–731.
26. Egelhofer, V.; Bussow, K.; Luebbert, C.; Lehrach, H.; Nordhoff, E. Improvements in Protein Identification by MALDI-TOF-MS Peptide Mapping. *Anal. Chem.* **2000,** *72(13),* 2741–2750.
27. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C.* Cambridge University Press: Cambridge, 1995.699–706.
28. Gusev, A. I.; Wilkinson, W. R.; Proctor, A.; Hercules, D. M. Improvement of Signal Reproducibility and Matrix/Comatrix Effects in MALDI Analysis. *Anal. Chem.* **1995,** *67(6),* 1034–1041.
29. Duncan, M. W.; Matanovic, G.; Cerpa-Poljak, A. Quantitative Analysis of Low Molecular Weight Compounds of Biological Interest by Matrix-Assisted Laser Desorption Ionization. *Rapid Commun. Mass Spectrom.* **1993,** *7(12),* 1090–1094.
30. Gobom, J.; Kraeuter, K. O.; Persson, R.; Steen, H.; Roepstorff, P.; Ekman, R. Detection and Quantification of Neurotensin in Human Brain Tissue by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Anal. Chem.* **2000,** *72(14),* 3320–3326.
31. Gusev, A. I.; Wilkinson, W. R.; Proctor, A.; Hercules, D. M. Direct Quantitative Analysis of Peptides Using Matrix Assisted Desorption Ionization. *Fresenius J. Anal. Chem.* **1996,** *354,* 455–463.
32. Kratzer, R.; Eckerskorn, C.; Karas, M.; Lottspeich, F. Suppression Effects in Enzymatic Peptide Ladder Sequencing Using Ultraviolet-Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry. *Electrophoresis* **1998,** *19(11),* 1910–1919.
33. Han, J.; Kamber, M. *Data Mining.* Academic Press: San Diego, 2001,354–362.
34. Medzihradszky, K. F.; Campbell, J. M.; Baldwin, M. A.; Falick, A. M.; Juhasz, P.; Vestal, M. L.; Burlingame, A. L. The Characteristics of Peptide Collision-Induced Dissociation Using a High-Performance MALDI-TOF/TOF Tandem Mass Spectrometer. *Anal. Chem.* **2000,** *72(3),* 552–558.
35. Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. A Tandem Quadrupole/Time-of-Flight Mass Spectrometer With a Matrix-Assisted Laser Desorption/Ionization Source: Design and Performance. *Rapid Commun. Mass Spectrom.* **2000,** *14(12),* 1047–1057.
36. Wei, J.; Buriak, J. M.; Siuzdak, G. Desorption-Ionization Mass Spectrometry on Porous Silicon. *Nature* **1999,** *399(6733),* 243–246.