Protein identification by MALDI-TOF-MS peptide mapping: a new strategy

ARTICLE *in* ANALYTICAL CHEMISTRY · MAY 2002

Impact Factor: 5.64 · DOI: 10.1021/ac011204g · Source: PubMed

6 AUTHORS, INCLUDING:

Johan Gobom
University of Gothenburg
76 PUBLICATIONS **3,143** CITATIONS
SEE PROFILE

Harald Seitz
Fraunhofer Institute for Cell Therapy and Im…
45 PUBLICATIONS **1,411** CITATIONS
SEE PROFILE

Hans Rudolf Lehrach
Max Planck Institute for Molecular Genetics
855 PUBLICATIONS **70,257** CITATIONS
SEE PROFILE

Anal. Chem. **2002,** *74,* 1760−1771

Protein Identification by MALDI-TOF-MS Peptide Mapping: A New Strategy

Volker Egelhofer,*,†,‡ Johan Gobom,† Harald Seitz,† Patrick Giavalisco,† Hans Lehrach, and Eckhard Nordhoff†,‡,§

Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany, and Protagen Ag, Im Lottental 36, 44801 Bochum, Germany

A new strategy for identifying proteins by MALDI-TOF-MS peptide mapping is reported. In contrast to current approaches, the strategy does not rely on a good relative or absolute mass accuracy as the criterion that discriminates false positive results. The protein sequence database is first searched for all proteins that match a minimum five of the submitted masses within the maximum expected relative errors when the default or externally determined calibration constants are used, for instance, ±500 ppm. Typically, this search retrieves many thousand candidate sequences. Assuming initially that each of these is the correct protein, the relative errors of the matching peptide masses are calculated for each candidate sequence. Linear regression analysis is then performed of the calculated relative errors as a function of *m*/*z* for each candidate sequence, and the standard deviation to the regression is used to distinguish the correct sequence among the candidates. We show that this parameter is independent of whether the mass spectrometric data were internally or externally calibrated. The result is a search engine that renders internal spectrum calibration unnecessary and adapts to the quality of the raw data without user interference. This is made possible by a dynamic scoring algorithm, which takes into account the number of matching peptide masses, the percentage of the protein's sequence covered by these peptides and, as new parameter, the determined standard deviation. The lower the standard deviation, the less cleavage peptides are required for identification and vice versa. Performance of the new strategy is demonstrated and discussed. All necessary computing has been implemented in a computer program, free access to which is provided in the Internet.

MALDI-TOF-MS peptide mapping is the most frequently used method for identifying proteins in sequence databases.[1−12] In brief, the method comprises the following steps: the protein to be identified is digested with a specific protease, most frequently trypsin, and the resulting peptides are mass analyzed. These data are compared with expected values computed from sequence database entries. The results are scored, and the value of the highest score and its difference to the score of the next following, nonrelated sequence suggest the protein being identified or not. If no significant restrictions can be put on the proteins' molecular weight or a large database is searched, a mass accuracy better than 50 ppm is required for identification.[13]

Because the calibration constants are sample position-dependent, state-of-the-art MALDI-TOF mass spectrometers meet the above requirements only if the recorded mass spectra are calibrated internally using a minimum two reference signals. A disadvantage of internal calibrants is that they affect the detection of the analyte molecules and vice versa due to competition for incorporation into, or adsorption to the matrix crystals as well as competition for charge during MALDI. In addition, analyte signals may overlap with the calibrant signals and thereby ruin the calibration or be excluded from the analysis. Externally determined calibration constants, on the other hand, can restrict the mass accuracy to 100 ppm or worse.

As previously shown, sample position-dependent mass errors are systematic in nature. This observation was made use of by a protein identification strategy, which corrects for these errors using information contained in the opened sequence database.[14] The strategy reported here does not require manipulating the experimental data and makes use of the observation that the

* To whom correspondence should be addressed: (phone) +49 30 6392 1749; (fax) +49 30 6392 1701; (e-mail) egelhofer@scienion.de.

† Max-Planck-Institute for Molecular Genetics.

‡ Current address: Scienion AG, Volmerstrasse 7b, 12489 Berlin, Germany.

§ Protagen Ag.

(1) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.

(2) Yates, J. R., III; Speicher, S.; Griffin, P.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.

(3) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *1*, 58−64.

(4) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.

(5) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−345.

(6) Cottrell, J. S. *Pept. Res.* **1994**, *7*, 115−124.

(7) Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791−1814.

(8) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.

(9) Hochstrasser, D. F. *Clin. Chem. Lab. Med.* **1998**, *36*, 825−836.

(10) Page, M. J.; Amess, B.; Rohlff, C.; Stubberfield, C.; Parekh, R. *Drug Discovery Today* **1999**, *4*, 55−62.

(11) Wilkins, M. R.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics*; Springer-Verlag, 1999.

(12) Blackstock, W. P.; Weir, M. P. *Tiptech* **1999**, *17*, 121−127.

(13) Jensen, O. N.; Podtelejnikov, A.; Mann, M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1371−1378.

(14) Egelhofer, V.; Büssow, K.; Luebbert, C.; Lehrach, H.; Nordhoff, E. *Anal. Chem.* **2000**, *72*, 2741−2750.

10.1021/ac011204g CCC: $22.00
Published on Web 03/16/2002

relative (and not the absolute) mass errors that result when moving from one sample position to an other correlate linearly with m/z.[15]

The strategy comprises three steps, all of which are performed by a new version of the database search program MSA,[14] free access to which will be provided at http://www.scienion.de/msa. First, the database is searched for all proteins that contain a minimum five cleavage peptides whose molecular masses match the determined masses within the maximum expected deviation according to the instrument's specifications for the use of the default or externally determined calibration constants, for instance, ±500 ppm. If all entries of a large database such as the NCBI sequence database are examined, this search typically retrieves many thousand candidate sequences. As second step, MSA calculates the relative errors of the matching peptide masses for each candidate, making the initial assumption that each candidate is the correct sequence. Linear regression analysis is performed on the relative errors for each candidate sequence as a function of m/z, and the standard deviation (SD) to the regression line for each candidate is then used to discriminate false positives. This approach is flexible in that it characterizes and adapts to the quality of the experimental data. In addition, it renders internal or close external spectrum calibration unnecessary.

The identification relies on a dynamic scoring algorithm that for each sequence candidate takes into account the following: the number of hits n, the SD, and the percentage of the protein's sequence covered by the matched peptides (sequence coverage, SC). As third and last step, MSA calculates the score Z for each sequence and ranks these accordingly. The smaller the SD and the larger the value of n and SC, the higher is the score. As a consequence, the lower the SD, the less cleavage peptides or less SC are required to identify the protein and vice versa. This approach renders fully automated database searches in batch mode straightforward.

Performance of the strategy was evaluated by the identification of several hundred different recombinant proteins from *Arabidopsis thaliana* and human, whose identity was independently confirmed by DNA sequencing. Examples are shown and discussed as well as examples of native proteins, separated by two-dimensional gel electrophoresis.

EXPERIMENTAL SECTION

Materials. The following three poly(propylene glycol) (PPG) oligomer fractions were purchased from Aldrich (Milwaukee, WI): M_n 1000 (20,232-0), M_n 2000 (20,233-9), and M_n 2700 (20,234-7). The following peptides (peptide 1−8) were purchased from Bachem: human angiotensins I and II, substance P-methyl ester, neurotensin, neurotensin (clip 1−11), ACTH (clip 1−17), ACTH (human clip 18−39), and somatostatin. α-Cyano-4-hydroxycinnamic acid (CHCA) was purchased from Sigma (St. Louis MO) and *n*-octylglucopyranoside (*n*-OGP) was obtained from Fluka.

Peptide Calibration Standard. A stock solution containing 1 pmol/μL of each of the peptides 1−6 and 2 pmol/μL of peptides 7 and 8 in 35% acetonitrile (v/v), 0.1% trifluoroacetic acid (TFA) (v/v) was prepared by following the quantity specifications provided by the manufacturer. This solution was diluted 20 times with a 5 mM solution of *n*-OGP in 0.1% TFA and used as the peptide calibration mixture. This standard was prepared for analysis as described.[16]

PPG Calibration Standard. The three PPG fractions were diluted 1:10 000 (v/v) in 99% acetone, 0.001% TFA (v/v). The three fractions were mixed in the ratio 1:2:3 (M_n 1000:2000:2700) (v/v), aliquoted, and stored at −20 °C prior to use. The MALDI matrix solution was prepared by ultrasonicating an excess of CHCA in 99% acetone, 0.001% TFA (v/v) for 1 min. The matrix solution was mixed with the PPG calibrant mixture at a ratio of 4:1 (v/v). A few sodium chloride crystals were added to the solution to enhance sodium cationization of the PPG molecules. A volume of the solution was aspirated by capillary force into a narrow pipet tip (GELoader, Eppendorf). MALDI samples of the calibrant mixture was prepared by touching the outlet of the pipet tip onto the hydrophilic sample anchors of a Scout MTP prestructured sample support (Scout 384-MTP AnchorChip, Bruker Daltonik, Bremen, Germany), whereby a small volume of the PPG/matrix solution was deposited.

Tryptic Digests of Recombinant Proteins. cDNA expression clones were selected from the human brain expression cDNA library hEx1.[17] Several hundred clones of this library were DNA sequenced. For all these clones, the encoded recombinant proteins were expressed, metal-affinity purified, and digested with trypsin as described.[18,14] Recombinant *A. thaliana* proteins expressed in *Escherichia coli* were kindly provided by Dr. Kersten, Max-Planck-Institute for Molecular Genetics, Berlin. The purified proteins were digested with trypsin as described.[14]

Tryptic Digests of Native Proteins. Coomassie G250-stained large-format 2D gels[19,20] of human brain total protein extract were a gift from Prof. J. Klose, Institute for Human Genetics, Virchow-Klinikum, Humboldt-University, Germany. Cylindrical gel samples of 1-mm diameter were excised and digested in situ as described.[21]

Peptide Sample Preparation for MALDI. All peptide samples were prepared using the CHCA surface affinity preparation, previously described.[16]

Mass Spectrometric Analyses, Data Processing, and Database Searching. All mass spectra were recorded automatically on a Bruker Scout MTP Reflex III mass spectrometer in reflector mode. Exclusively positively charged ions were analyzed in the reflector mode, and 200 single-shot spectra were accumulated for improved signal-to-noise ratio. The extraction delay time was 150 ns, and deflection was used to suppress ions up to m/z 500. The data were sampled with 2 GHz. All further processing except spectrum calibration was performed in batch mode using the software package XMASS 5.1, provided by the instrument manufacturer. The raw sum spectra were baseline corrected before automated peak picking was performed using

(15) Gobom, J.; Egelhofer, V.; Mueller, M.; Lehrach, H.; Nordhoff, E. *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, Chicago, IL, 2001; ThPA 001.

(16) Gobom, J.; Schuerenberg, M.; Mueller, M.; Theiss, D.; Lehrach, H.; Nordhoff, E. *Anal. Chem.* **2001**, *73*, 434−438.

(17) Büssow, K.; Cahill, D.; Nietfeld, W.; Bancroft, D.; Scherzinger, E.; Lehrach, H.; Walter, G. *Nucleic Acids Res.* **1998**, *26*, 5007−5008.

(18) Büssow, K.; Nordhoff, E.; Luebbert, C.; Lehrach, H.; Walter, G. *Genomics* **2000**, *65*, 1−8.

(19) Klose, J.; Kobalz, U. *Electrophoresis* **1995**, *16*, 1034−1059.

(20) Klose, J. Large-Gel 2-D Electrophoresis. In *2-D Proteome Analysis Protocols*; Link, A. J., Ed.; Methods in Molecular Biology 112; Humana Press Inc.: Totowa, NJ, 1999; pp 147−172.

(21) Nordhoff, E.; Eickhoff, H.; Horn, M.; Przewieslik, T.; Egelhofer, V.; Giavalisco, P.; Theiss, D.; Lehrach, H.; Gobom J. *Electrophoresis* **2001**, *22*, 2844−2855.

the algorithm SNAP provided by XMASS 5.1. This algorithm uses the data points for all recorded monoisotopic signals of a peptide to assign a flight time or, if a calibration is applied, an m/z value to the first monoisotopic peak. No peak filtering was performed prior to database searching, and no restrictions were put on the molecular weight of the proteins. Unless otherwise stated, the number of possible missing cleavage sites was restricted to one, all cysteine residues were expected to be reduced, and variable modification were not considered.

Calibration of the Recorded Time-of-Flight Spectra. For external calibration, the two calibration standards (see above) were used. When the peptide calibration standard was used, the first monoisotopic signals assigned to the eight peptides were labeled in the calibrant spectrum. Based on the square of the determined flight times and the corresponding calculated m/z values, a linear calibration function was established using the software routine provided by XMASS 5.1. This calibration was then applied to convert peptide time-of-flight spectra recorded on other positions of the same support to mass spectra. When the PPG calibration standard was used, the 58 sodium-cationized PPG molecular ions in the range m/z 737−4046 were labeled in the calibrant time-of-flight spectrum. Instead of a linear calibration, the Householder curve-fitting algorithm implemented in LabView 6.0 (National Instruments) was used to determine the coefficients in a 15-order polynomial function for the relationship between the calculated monoisotopic masses of the PPG sodium cations and the square of their assigned flight times. This function was then used to convert the square of the determined flight times of ions detected in other samples to m/z values.

RESULTS AND DISCUSSION

The identification strategy reported here relies on the observation that, in MALDI-TOF-MS, the relative mass errors that result when moving from one sample position to an other correlate linearly with m/z.[15] Figure 1 shows four examples that confirm this observation. Four equal aliquots of a tryptic digest of the *A. thaliana* transcription factor TGA3 (NCBI, 1076421), expressed in *E. coli*, were prepared on a Bruker Scout 384/400 AnchorChip sample support on positions G11, G15, I11, and I15 (Figure 1a) using the CHCA-affinity sample preparation technique introduced recently[16] and analyzed in automatic acquisition mode. All assigned flight times (flight time peak lists) were converted to m/z values using the same set of calibration constants, which before were determined for the PPG calibration standard prepared on position G10. The recorded mass spectrum of the PPG standard is shown in Figure 1b, and the spectra of the four peptide samples are reproduced in Figure 1c−f. Signals matching tryptic peptides of the transcription factor are green. Red signals could not be assigned, and the blue signals indicate satellite signals that represent oxidized methionine and tryptophan residues (see below: Variable Modifications of Specific Amino Acid Residues). A comparison by eye shows that the four spectra are, as expected, very similar. The relative errors of the assigned peptide masses, plotted underneath the spectra as a function of m/z, however, varied significantly from one spectrum to the next. At position G11, next to the calibration standard, the maximum relative error was less than −25 ppm, at position G15 it was close to −50 ppm, and at position I15 it exceeded −100 ppm. In contrast to this, the SD of the relative mass deviations to the regression lines for each plot changed little (maximum deviation, 0.3 ppm). For positions G11, G15, I11, and I15 it was 4.3, 4.1, 4.0, and 4.0 ppm, respectively. These data confirm the observation that position-dependent relative mass deviations correlate linearly with m/z.

On the basis of experiments with PPG and peptide standards prepared on all 384 positions of the sample support, it was established empirically that, with the instrumentation used, the slope of the regression line for samples on any position on the MALDI sample support never exceeded 0.03 ppm/(m/z). In many cases, the slope was negligible, which means that in these cases also the absolute errors were directly proportional to m/z. It was also determined that no relative error across the entire sample support was larger 300 ppm when the same calibration constants were used for all samples. A comparison of the used AnchorChip sample supports versus unmodified stainless steel support showed no difference with respect to the characteristics of the relative error plots discussed above.

Proposed Strategy. The proposed strategy comprises three steps and is outlined in Figures 2 and 3. A tryptic digest of recombinant human β-actin (cytoplasmic 1) was analyzed on position G16, and the recorded flight times were converted to m/z using an external two-point calibration, established for the peptide calibration standard on position G12. Figure 3a shows the recorded spectrum of the sample. The peak-picking algorithm generated a list of 37 entries which, after conversion from m/z to molecular masses, was submitted to MSA to retrieve all protein sequences contained in the NCBI database (release: October 21, 2001 containing >700 000 entries) that contained minimum five tryptic peptides with masses matching entries in the submitted peak list within ±500 ppm. Other conditions were as follows: one allowed missing cleavage, all cysteine residues reduced, and no other modifications expected. The search retrieved 23 991 sequences. As expected, the correct sequence was among these, matched by 21 out of 37 submitted masses.

The second step starts with the assumption that all peptide masses contained in the 23 991 mass lists computed in step one are true hits. The program MSA treats these and the underlying sequences equally and applies to each data set a series of simple calculations. In the following example, the calculations are outlined for the correct sequence of human β-actin. First, for the 21 matched peptide masses the mean, μ, and the standard deviation to the mean, SD_μ, of their relative deviations to the calculated values were calculated to −144.9 and +180.7 ppm, respectively. The calculated relative deviations as a function of m/z are plotted in Figure 3b and range from −299 to +430 ppm. The two dashed red lines at −506 and +217 ppm indicate $\mu - 2SD_\mu$ and $\mu + 2SD_\mu$. Matching peptide masses with relative mass errors outside this interval were assigned as outliers and excluded from the following calculations. In the current example, two data pairs (indicated in red in the plot) exceeded the upper limit. Hereafter, a linear fit, described by the function $Y' = 0.0218X - 238.2$ (the regression line is shown in Figure 3c), was calculated for the remaining 19 data pairs. The standard deviation to Y' ($SD_{Y'}$) was calculated to 40.5 ppm. In Figure 3d, instead of the relative mass deviations, the relative deviations to Y' are plotted as a function of m/z. Analogously to the first step, the standard deviation was used to remove outliers. In the current example, two of the remaining 19 hits fell outside the limits $-2\ SD_{Y'}$ and $+2\ SD_{Y'}$, calculated to −81

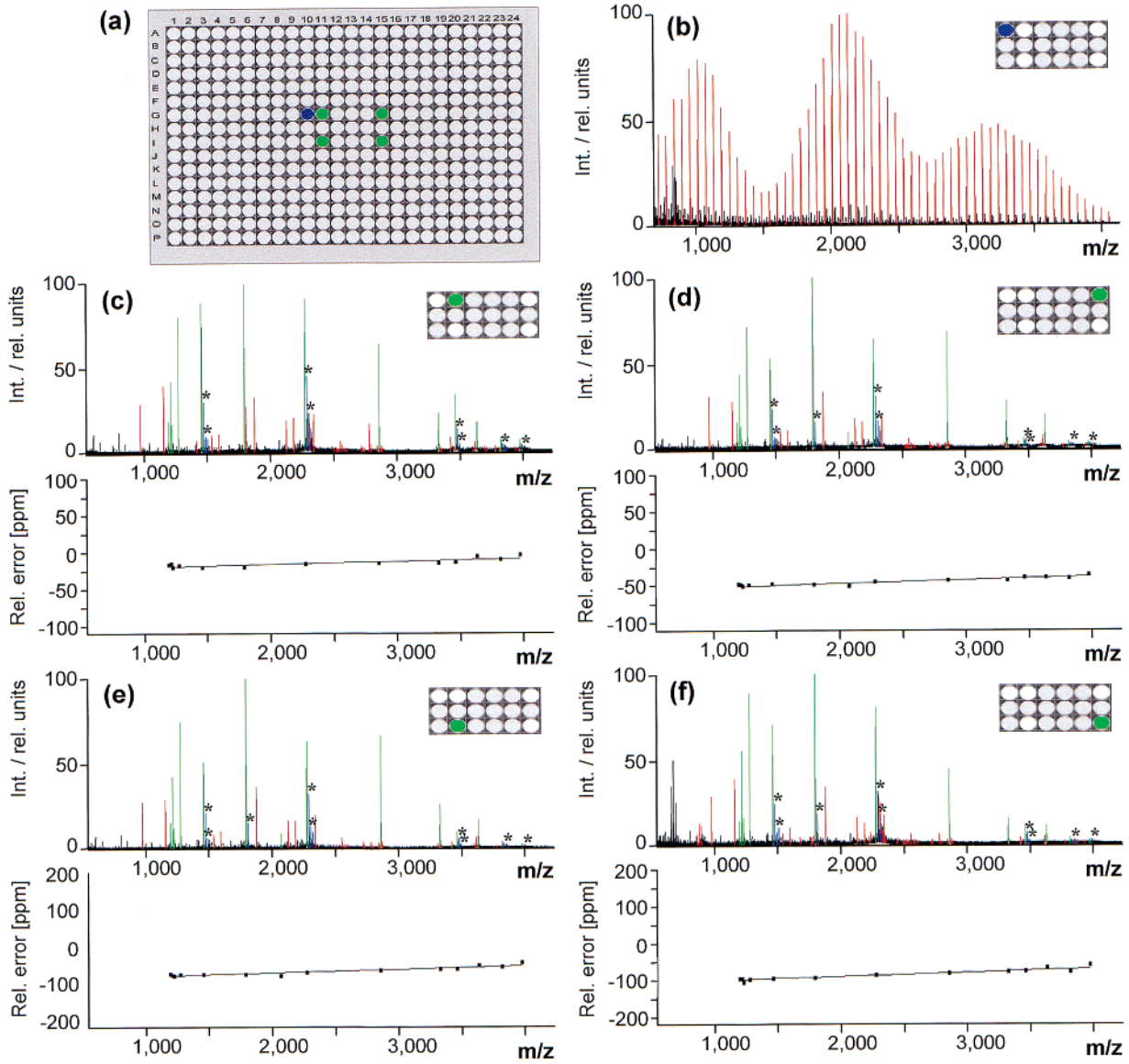

Figure 1. Linear correlation of position-dependent relative mass deviations with *m*/*z* in MALDI-TOF-MS. (a) Geometry of the MALDI sample support and the sample positions used in this experiment: blue, PPG standard; green, a tryptic digest of recombinant *A. thaliana* transcription factor TGA3 expressed in *E. coli.* (b) Recorded mass spectrum of the PPG standard: red, signals assigned to PPG molecular ions. (c−f) Mass spectra recorded on positions G11, G15, I11, and I15: green, signals assigned to tryptic peptides of TGA3; red, analyzed signals that could not be assigned; signals that are blue and marked with a star indicate oxidation of methionine or tryptophan residues. Below: the relative error plot of the determined peptide masses assigned to TGA3. The trend line was calculated by a linear regression. For the four data sets, the standard deviation to the fit varied by maximum 0.3 ppm. At positions G11, G15, I11, and I15, it was calculated to 4.3, 4.1, 4.0, and 4.0 ppm, respectively.

and +81 ppm and marked in the figure by two dashed red lines. After exclusion of these outliers, the linear fit was updated ($Y = 0.0232X - 239.6$) according to the remaining values (Figure 3e), with the result that the standard deviation SD_Y dropped from 40.5 to 13.7 ppm (Figure 3f). This value is independent of the intercept but depends on the slope of the calculated regression. Thus, larger values would ensue, the more the slope deviated from zero. This is overcome according to

$$SD = SD_Y/(1 + \text{slope}^2)^{0.5} \qquad (1)$$

where SD represents the standard deviation of the relative errors not described by *Y*. While the intercept and the slope of *Y* account for the systematic, sample position-dependent part of the mass errors, the remaining part is characterized by SD. For scoring the candidate sequences, this value replaces mass accuracy as the important parameter to discriminate false positive results. In the example of β-actin, the correction factor $(1 + \text{slope}^2)^{0.5}$ was calculated to 0.9998 and had no influence on the identification (SD = 13.7 ppm = SD_Y). For the maximum slope of 0.03 observed in our laboratory (see above), the factor is 0.9995. It appears that above transformation of SD_Y to SD is unnecessary, considering that only the integer value and the first decimal of SD are considered in the following calculations. Whether this conclusion is generally true or not, however, is not clear because said slope might be affected by the instrument design as well as the sample preparation (initial velocity of the generated ions). Therefore, we

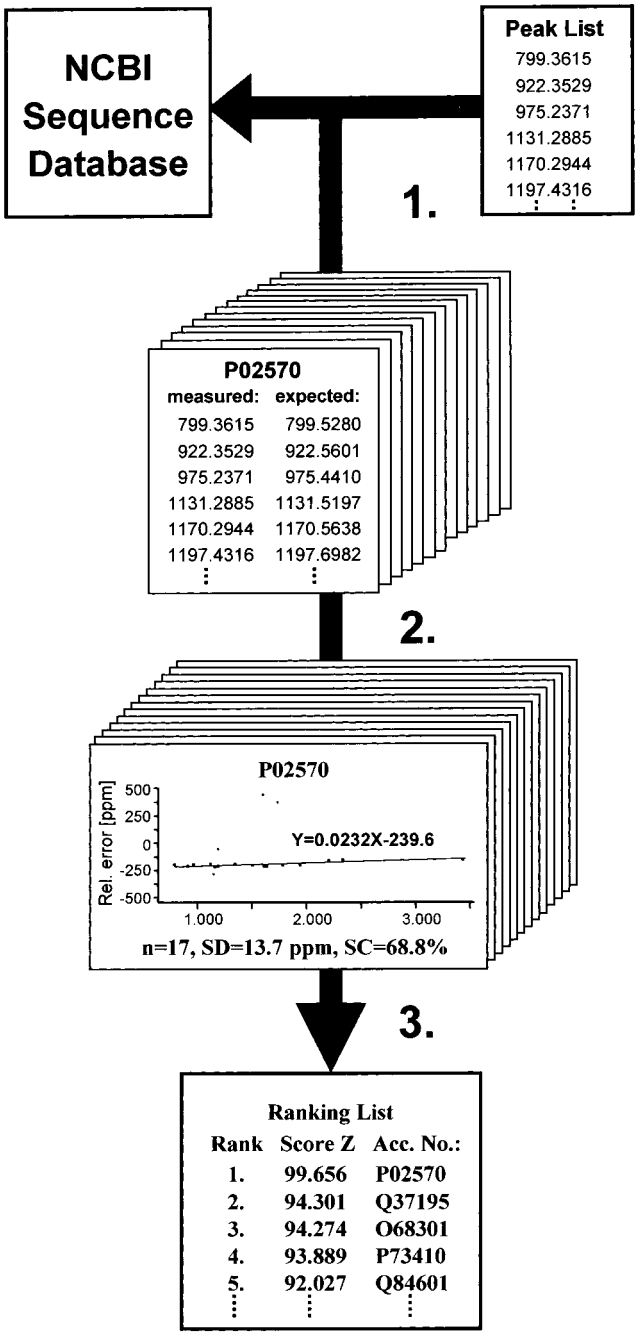

Figure 2. Proposed protein identification strategy. The strategy is based on MALDI-TOF-MS peptide mapping and comprises three steps. Step 1: the database is searched for all protein sequences that match a minimum five of the submitted masses within the maximum expected relative error when the default or externally determined calibration constants are used, for instance, ±500 ppm. Assuming initially that each of these is the correct sequence, the relative errors of all matching peptide masses are calculated. Step 2: For each candidate sequence, a linear regression analysis is performed of the calculated relative errors as a function of *m*/*z*, which also includes recognition of outliers (error > 2SD, marked red in the plot). For the corrected regression, the SD is calculated as well as the percentage of the protein sequence (SC) covered by the remaining hits (*n*). Step 3: Based on *n*, SD, and SC, for each candidate sequence, the *Z* score is calculated according to eq 2, and the results are ranked accordingly.

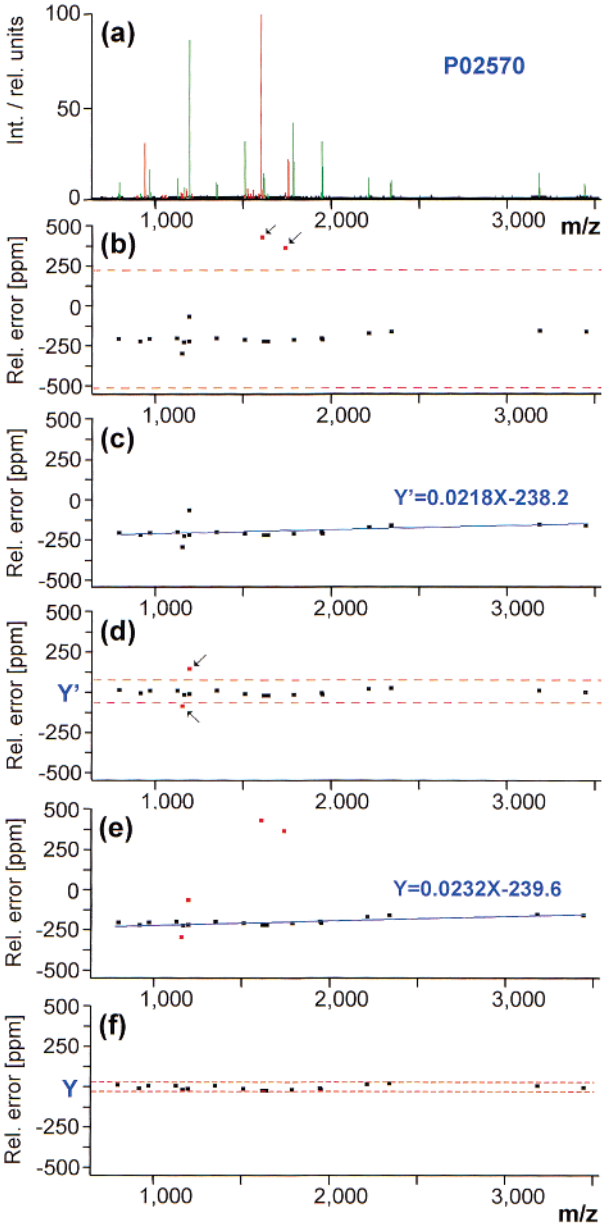

Figure 3. Example of the regression analysis used for protein identification. (a) MALDI-TOF mass spectrum obtained from a tryptic digest of recombinant human β-actin (cytoplasmic 1). (b) Relative error plot of the 21 out of 37 determined masses that matched tryptic peptides of the correct protein sequence, retrieved from the NCBI database, with a maximum relative error of ±500 ppm. The dashed red lines indicate the limits $\mu - 2SD_\mu$ and $\mu + 2SD_\mu$, which exclude outliers from the following calculations. (c) Trend line of the linear regression, *Y*′, calculated for the remaining 19 hits. (d) Relative deviations to *Y*′ plotted as a function of *m*/*z*. The standard deviation to *Y*′ ($SD_{Y'}$) was calculated to 40.5 ppm. The dashed lines indicate the limits $-2SD_{Y'}$ and $+2SD_{Y'}$ used to exclude two additional outliers from the following calculations. (e) Trend line of the linear regression, *Y*, calculated for the remaining 17 hits. (f) Relative deviations to *Y* plotted as a function of *m*/*z*. The standard deviation to *Y* (SD_Y) was calculated to 13.7 ppm.

select the parameter SD, which is independent of the slope of the calculated linear regression. In addition to SD, the percentage of the retrieved sequence of β-actin covered by the remaining 17 hits (SC) was calculated to 68.8%.

The above calculations including the removal of outliers were applied to all 23 991 sequences retrieved from the database, and for each of them, the three values *n* (number of valid hits), SD,

and SC were calculated. These numbers provide the data based on which, as the third part of the strategy, the score Z is calculated for each sequence according to

$$Z = 100 - (F \times 500\mathrm{SD})/(n^2\mathrm{SC}) \qquad (2)$$

Z determines whether a specific sequence is considered a false positive, a possible candidate, or very likely to be a true hit. F is an arbitrary factor that allows modification of the stringency of the search. The default value for F, exclusively used in this study, is 1.0. Smaller values increase and larger values decrease the value of Z, with the consequence that the conditions for identification (see below) become less or more stringent.

The scoring function has been designed such that Z approaches but never reaches the value 100, which would imply 100% certainty, an impossible result. Sequences that yield a score equal to or greater than 99 are considered likely to be true positives, values equal or above 98 and below 99 indicate possible but uncertain candidates, and values below 98 are considered insufficient for identification.

For the correct sequence of human β-actin, the score was calculated to 99.656, a value that suggests the protein is identified. In this case, more than 50 sequence entries yielded the same or a very similar score. These, however, where all entries for β-actin or closely related homologues, and in all cases, the same peptide sequences were matched. Not a surprising result, considering that the NCBI database contains several hundred actin sequences and no restrictions were put on the selection of species. More importantly, no nonrelated sequence (false positive) was ranked higher than 97.5.

An example for a true false identification is shown in Figure 4a. A tryptic digest of recombinant human PDZ domain protein 3′ (variant 4) expressed in *E. coli* was analyzed, and the generated peak list, after conversion from m/z to mass, was used to search all mammalian protein sequences contained in the SwissProt database (release, October 2001). The other conditions were identical to those named above. The identity of the protein had been independently confirmed by DNA sequencing, and it was found that its sequence was not contained in the SwissProt database. Consequently, any identification suggested by the search result ($Z \geq 99$) would have been a false result. The generated ranking list was headed by human clathrin heavy chain 2 (accession No. P53675) with a Z score of only 93.365 ($n = 13$, SD = 29.2 ppm, SC = 13.0%), far too low to suggest this sequence was identified or to be a possible candidate. If the entire NCBI database was searched instead, however, the correct sequence (NCBI, 11933155) was identified with a score of 99.363 ($n = 18$, SD = 10.9 ppm, SC = 26.4%). For comparison, the matched signals and the respective relative error plot are shown in Figure 4b.

Recognition and Removal of Outliers. For human β-actin, four of the submitted masses were assigned outliers. The relative error plots shown in Figure 3b and d confirm that these decisions were correct. In the calculation of Z, their loss was compensated for by the resulting drastic reduction of the SD. The two-step approach to recognize and remove outliers was necessary because the initial search was performed with a high error tolerance (±500 ppm). In this case, for instance, absolute errors of 1 Da are accepted above m/z 2000. Compared to other errors, absolute

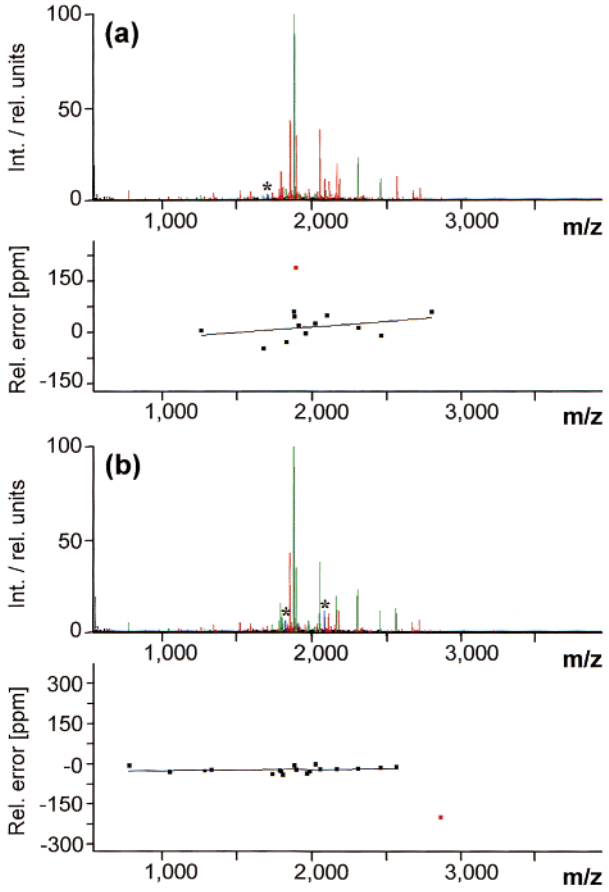

Figure 4. Example of a true negative and a true positive result. A tryptic digest of human PDZ domain protein 3′ (variant 4) recombinantly expressed in *E. coli* was analyzed by MALDI-TOF-MS, and the generated peak list was used to search all mammalian protein sequences contained in the SwissProt and in the NCBI database (release, October 2001). The identity of the protein had previously been verified by DNA sequencing, and it was found that its sequence is contained in the NCBI but not in the SwissProt database (release, October 2001). As a correct result, no protein was identified in the SwissProt database, and the ranking list was headed by Clathrin heavy chain 2 (P53675) with a Z score of only 93.365. (a, b) The recorded mass spectrum with the signals assigned to Clathrin heavy chain (a) and the PDZ domain protein (b) are green and those that were not assigned are red; blue signals marked with a star indicate oxidation of methionine or tryptophan residues. Below the spectra the corresponding relative error plots are shown.

mass deviations to the mean around 1 Da are systematic and abundant, especially when peaks with a very poor signal-to-noise-ratio are assigned in the spectrum, as is often the case when very small amounts of proteins are to be identified. Under such conditions, the signals of the first monoisotopic peptide ions may not be distinguishable from the baseline noise and instead the signal of the more abundant species containing one ^{13}C atom is accidentally assigned. An example for this is shown in Figure 5, documenting identification of a mitochondrial precursor of human aconitate hydratase, isolated from a crude human brain protein extract by large-gel two-dimensional gel electrophoresis.[19,20] Six of 26 peptide masses that initially matched this protein with a maximum error of ±500 ppm were recognized as outliers and removed in step two of the identification. The remaining 20 hits yielded a Z score of 99.385 (SD = 21.9 ppm, SC = 44.5%), suggesting the protein was identified. In this example, as docu-

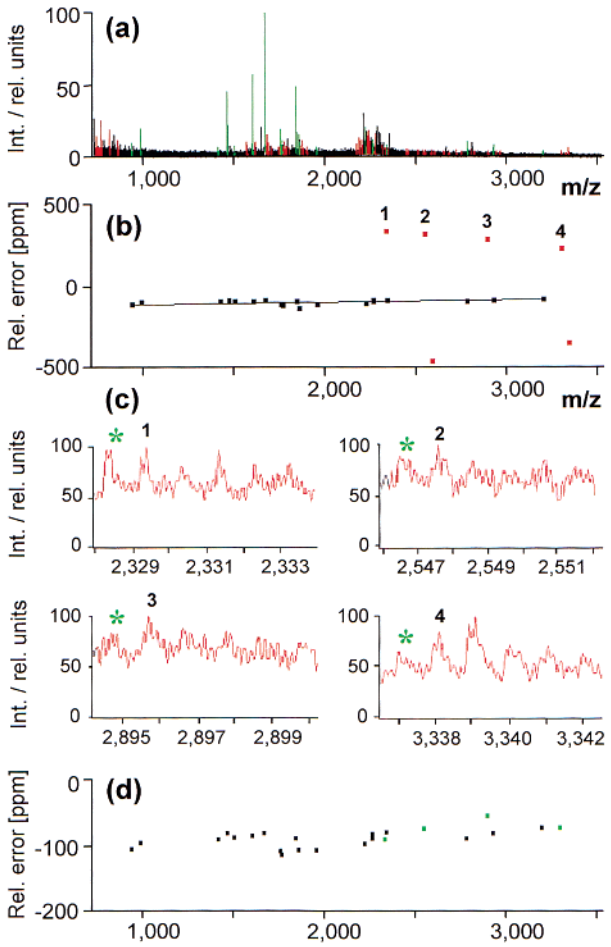

Figure 5. Example of an identification based on mass spectrometric data of poor quality. (a) MALDI-TOF mass spectrum obtained from a tryptic digest of human aconitate hydratase (mitochondrial precursor) isolated from a crude human brain protein extract by two-dimensional gel electrophoresis: green, signals assigned to tryptic peptides; red, signals that could not be assigned. (b) Relative error plot of the 26 determined masses that matched the correct protein sequence in the NCBI database with a maximum relative error of 500 ppm. During the identification process, six of these masses were assigned outliers; red squares in the plot. The remaining 20 hits with a SD of 21.9 ppm yielded a *Z* score of 99.385. (c) Inspection of the signals of the outliers revealed that, in four cases, due to poor signal-to-noise ratio, the peak-picking algorithm had assigned the second instead of first monoisotopic molecular ion signal; marked by a green star in the magnified *m*/*z* windows shown. (d) After correction, the identification was confirmed by four additional hits, shown in green in the relative error plot.

mented in Figure 5c, due to the poor quality of the raw data, the correct monoisotopic peak was not recognized in four cases and instead the second peak was assigned. Once recognized, the peak assignment could be corrected, with the result that four additional hits confirmed the identification (Figure 5d).

Scoring Algorithm. The scoring algorithm was empirically developed and tuned, based on the results obtained for several hundred recombinant mammalian and plant cDNA expression products, whose identity was independently confirmed by DNA sequencing. Table 1 provides a representative selection of these results (20 proteins) including the respective *Z* scores calculated according to eq 2. How this score responds to variations of SD in dependence of *n* for two different values of SC is plotted in Figure 6. The plots show that the score with increasing number of matched peptides approaches the limit 100 and how this is affected by the SD (2, 5, 10, 15, or 20 ppm). For the influence of the SC, the two cases 20% and 50% coverage are compared. *Z* increases with increasing SC and decreases with SD. As a consequence, the value of SD determines how many peptides need to be matched and how good the sequence coverage must be to identify any given protein. The value of SD depends on many different factors, e.g., the applied calibration method (see below), tuning of the mass spectrometric hardware, the data sampling rate, and the signal-to-noise-ratio of the assigned signals. If the latter is very poor, the nonsystematic part of the relative errors will generally be higher, with the consequence that the SD is higher. An example of this is shown in Figure 5. In contrast to the common practice that the user restricts the maximum expected absolute or relative error to a fixed value (e.g., 30 ppm), this method is more flexible because it adapts to the "quality" of the mass spectrometric data. Four additional examples document this in Figure 7.

Variable Modifications of Specific Amino Acid Residues. In addition to expected quantitative modifications of specific amino acid residues, e.g., after carbamidomethylation of cysteine residues, variable and partial modifications have to be considered when proteins are identified by MALDI-MS peptide mapping. Prominent examples are acrylamidation of cysteine residues and oxidation of methionine and tryptophan residues. As is common practice in other search programs (e.g., Profound, Mascot, MS-Fit), the current version of MSA takes both categories into account. Typically, variable and partial modifications are observed as satellite signals of the signal of the unmodified peptide. For instance, the mass spectra of a tryptic digest of the transcription factor TGA3 shown in Figure 1 contain five signals, of which four (at m/z 1459.7, 2277.0, 3819.7, and 3975.8) are followed by three and one (at m/z 3464.7) is followed by two satellite signals with a constant difference of m/z 16 (blue signals in the figure). The intensity of the satellite signals was lower than the intensity of the signal of the unmodified peptide, and the first satellite was more intense than the following. Inspection of the amino acid sequence of the matched peptides showed that each of the four for which three satellite signals were observed (amino acids 225−236, 18−36, 150−183, and 150−184, respectively) contains two methionine and one tryptophan residue. The peptide for which two satellites were observed (amino acids 308−339) contains two methionine residues. It appears straightforward to conclude that these hits are less likely to be false than those based only on one determined mass and that this information should be considered in the database search. The problem is that the abundance of the amino acid residue, whose presence was confirmed by satellite signals, can vary considerably from protein to protein and across different species. We have not found a general strategy for how assigned variable modifications should be considered in the scoring of the search results. Instead, our search program provides advanced settings, which allow the user to specify empirically for each expected variable modification how its assignment to a matched peptide affects the *Z* score. This information, however, is only considered in the calculation of *Z* (step three) and not for the selection of candidate sequences (step one) or the calculation of SD (step two).

Table 1. Summary of the Search Results Obtained for 20 Different Human cDNA Expression Products[a]

sample no.	rank	gene description	SwissProt+ acc	MW	hits	SD	SC	*n*	*Z*
	1	triosephosphate isomerase	p00938	26.538	10	9.9	59.3	68	99.165
	2	peripherin	P41219	53.878	6	6.5	19.3	68	0.000
2	2	heat shock cognate 71-kDa protein	P11142	70.898	22	7.5	56.8	115	99.864
	2	myosin heavy chain, skeletal muscle, perinatal	P13535	222.762	18	21.2	14.7	115	97.774
3	1	eukaryotic translation initiation factor 3 subunit 3	O15372	39.930	11	9	46.3	30	99.197
	2	nucleoprotein TPR	P12270	265.600	6	15.8	3.2	30	31.424
4	1	60S ribosomal protein L7A	P11518	29.864	5	1.5	33.6	81	99.107
	2	T-complex protein 1, θ subunit	P50990	59.620	6	7	13.1	81	92.578
5	1	mRNA-associated protein MRNP 41	P78406	40.968	11	1.9	45.4	48	99.827
	2	zinc finger protein 85	Q03923	68.718	6	11.9	16.6	48	90.044
6	1	proteasome subunit α type 1	P25786	29.555	9	5.5	57.4	63	99.409
	2	DNA damage binding protein 1	Q16531	126.967	6	6.2	8	63	89.236
7	1	mitotic centromere-associated kinensin	Q99661	81.312	14	11.8	35.9	87	99.162
	2	golgi autoantigen, golgin subfamily A 4	Q13439	261.139	18	29	11.1	87	95.968
8	1	bleomycin hydrolase	Q13867	52.562	14	2.2	49	30	99.885
	2	acetyl-CoA carboxylase 1	Q13085	265.038	10	29.6	8.1	30	81.728
9	1	glyceraldehyde 3-phosphate dehydrogenase, liver	P04406	35.922	12	3.7	53.6	46	99.760
	2	zinc finger protein 215	Q9UL58	60.048	6	18.3	17	46	85.049
10	1	kinesin light chain 1	Q07866	64.786	16	11.7	35.9	90	99.363
	2	serine/threonine protein phosphatase with EF-hands-2	O14830	86.430	7	11.6	15.7	90	92.461
11	1	guanylyl cyclase activating protein 1	P43080	22.774	8	7.4	68	47	99.150
	2	glutaminase, kidney isoform, mitochondrial precursor	O94925	73.461	6	9	19.9	47	93.719
12	1	elongation factor 1-γ	P26641	50.118	12	3.7	52.9	61	99.757
	2	voltage-gated potassium channel protein kV1.4	P22459	73.288	6	7.6	14.4	61	92.670
13	1	drebrin E	Q16643	71.425	9	4.9	29.1	51	98.961
	2	wolframin	O76024	100.305	7	12.6	15.5	51	91.705
14	1	40S ribosomal protein S4, X isoform	P12750	29.466	14	9.6	71.4	76	99.657
	2	UDP-glucuronosyltransferase 2B7 precursor, microsomal	P16662	60.694	6	2.7	15.3	76	97.549
15	1	vimentin	P08670	53.554	11	7.2	30.5	25	99.025
	2	DNA replication licensing factor MCM7	P33993	81.280	5	9.8	7.6	25	74.211
16	1	creatine kinase, B chain	P12277	42.644	13	3.6	61.7	32	99.827
	2	synaptojanin 2	O15056	159.953	6	13.2	6.7	32	72.637
17	1	ADP-ribosylation factor 3	P16587	20.469	10	8.7	78.9	79	99.449
	5	phenylalanine 4-hydroxylase	P00439	51.862	6	14	25.7	79	92.434
18	1	zinc-finger protein RFP	P14373	58.489	8	2.3	22.4	86	99.198
	2	calcium/calmodulin-dependent 3′,5′-cyclic nucleotide phosphodiesterase	P54750	61.120	5	9.5	22.7	86	91.630
19	1	probable ATP-dependent RNA helicase P47	Q13838	48.991	12	6	47.2	63	99.559
	2	calpain 3 large subunit	P20807	94.253	7	15.7	15.8	63	89.861
20	1	chromatin assembly factor 1 P48 l	Q09028	47.655	12	9.2	42.4	22	99.247
	2	cation-independent mannose 6-phosphate receptor precursor	P11717	274.306	6	15.4	5.3	22	59.644

[a] All proteins were identified in the NCBI and in the SwissProt database without species restriction, and in each case, the identification was independently confirmed by DNA sequencing. Search details are listed for the correct sequence and, underneath, for the second highest ranking, nonrelated protein sequence.

For the oxidation of methionine and tryptophan residues, frequently observed in our spectra, we found the following settings to yield good results. For each matched peptide, one assigned modification (satellite signal) counts as additional 0.6 hit, a second counts as 0.4 hit, and a third, the maximum accepted, counts as 0.25 hit. For instance, for the data reproduced in Figure 1e, compared to the default settings (no modifications expected), these settings increased the value for *n* from before 14 to 19 with the consequence that the score raised from before 99 873 to 99.931.

Influence of the Calibration Method. For the four mass spectrometric peptide maps shown in Figure 1a−d, the score *Z* for the correct sequence was calculated to 99.829, 99.870, 99.873, and 99.863, respectively. As expected, the value of *Z* was only little dependent on the sample position. However, compared to the example documented in Figure 3, the values calculated for SD were significantly smaller (4.3/4.1/4.0/4.0 versus 13.7 ppm). The main reason for this difference is that two different calibration methods were applied. In the first case (Figure 3), the determined flight times were converted to m/z using an external, higher-order multipoint calibration optimized for the spectrum of the PPG standard shown in Figure 1b.[15] For this purpose, a 15th-order polynomial regression was calculated that fits the square of the determined flight times to the calculated m/z values of the 58 PPG molecular ions spanning the m/z range 700−4000 (Figure 1b).

For the peptide map of human β-actin shown in Figure 3a, the determined flight times were converted to m/z by applying a

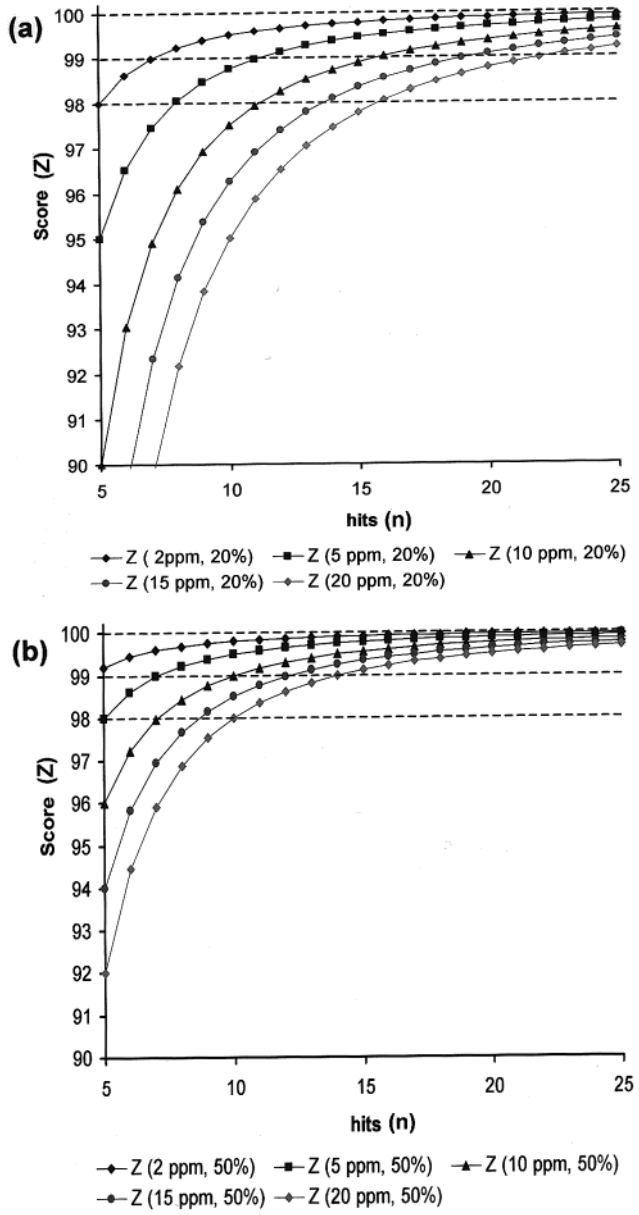

Figure 6. Response of the *Z* score to variations of the calculated SD in dependence of the number of hits, n. To visualize the influence of the achieved SC, the two cases (a) 20% and (b) 50% coverage are compared.

conventional linear calibration calculated for the peptide calibration standard. As control experiment (data not shown), the two calibration routines were exchanged, with the result that for the four examples shown in Figure 1, the value of SD increased to 12.6 ppm and for human β-actin it dropped to 4.8 ppm. We conclude that an optimized calibration can improve the performance of the proposed protein identification strategy. Examples that support this are provided in Figure 7. Whether this is necessary or not, depends on the performance of the mass spectrometric instrumentation, the acquisition parameters, and the size of the database to be searched. Although not suited for internal calibration, due to its complexity as well as a different ionization mode (sodium cationization versus protonation), the PPG calibration standard is well suited to externally calibrate peptide mass maps.[15] It is inexpensive, easy to store, and chemically stable, yields reproducible results, and therefore is well suited for automated spectrum acquisition. In addition, it is well suited to detect small instrumental changes that affect the correlation of flight time to m/z, because the correlation can be plotted over the full peptide mass range 700−4000 Da with a resolution of 58 Da. Besides the applied calibration method, the signal-to-noise-ratio is an important factor that determines the performance of MALDI-TOF-MS peptide mapping. The accuracy of the assigned flight times deteriorates when the detection sensitivity is approached, resulting in broader error distributions. That the developed scoring algorithm adapts to this situation is demonstrated in Figures 5 and 7a.

It is clear that the mass accuracy achievable by internal calibration depends on the number of calibrants available, their mass, and the quality of the corresponding signals. The relative error plots shown in Figure 7a−d visualize the underlying problem. Whether the maximum relative error, if any two of the assigned peptide signals were used for internal calibration, compared to each other is higher or lower depends on the choice of the two. Only those that fall close to the regression line are good candidates. This problem is circumvented by the proposed strategy. When relying on internal spectrum calibration, an even more severe problem arises when the necessary reference signals are absent. Frequently, signals of autoproteolysis products of the used protease fulfill this task. Possible reasons for the absence of these signals are low autoproteolysis rates or suppression effects in the sample preparation (competition for binding sites) or the subsequent analysis (competition for charge). This problem was another motivation to develop the strategy reported here.

Influence of the Sequence Coverage. The SC depends on the size of the protein, the number of matched peptides, their size, and whether their sequences overlap or not. If they overlap, the SC is lower, with the consequence that the value of *Z* decreases. If an additionally matched peptide is large (e.g., >3000 Da) and overlaps little or not with the other, its impact on the SC, and thereby on the *Z* score, will be significant. If it is small (e.g., < 1000 Da) and its sequence is already partly covered, its impact will be low. The calculation of *Z* requires that the larger a protein the more hits are required for identification. A constraint toward large proteins appears reasonable taking into account that the more peptide masses are predicted, the higher the risk for false positive hits. Our approach, although realized differently, follows the strategy proposed by Pappin et al.[4]

Influence of the Number of Submitted Masses. In addition to the large number of possible proteolytic peptides contained in long protein sequences, the number of submitted masses increases the risk for false positive identifications. The more entries in the peak list, the higher the risk for false positive hits. To evaluate whether the performance of our identification strategy is sensitive to the number of submitted peptide masses, automatic peak assignment using the SNAP algorithm was performed twice on the mass spectra acquired from the tryptic digests of 20 different recombinant *A. thaliana* proteins, once using our standard parameters, and once applying no limits to the signal detection sensitivity and a very low limit to the quality of the peaks. Both peak lists were then used to search all plant protein sequences contained in the NCBI database. In all cases, both peak lists identified the correct protein sequence. When no restrictions were applied to the peak-picking sensitivity, however, in most

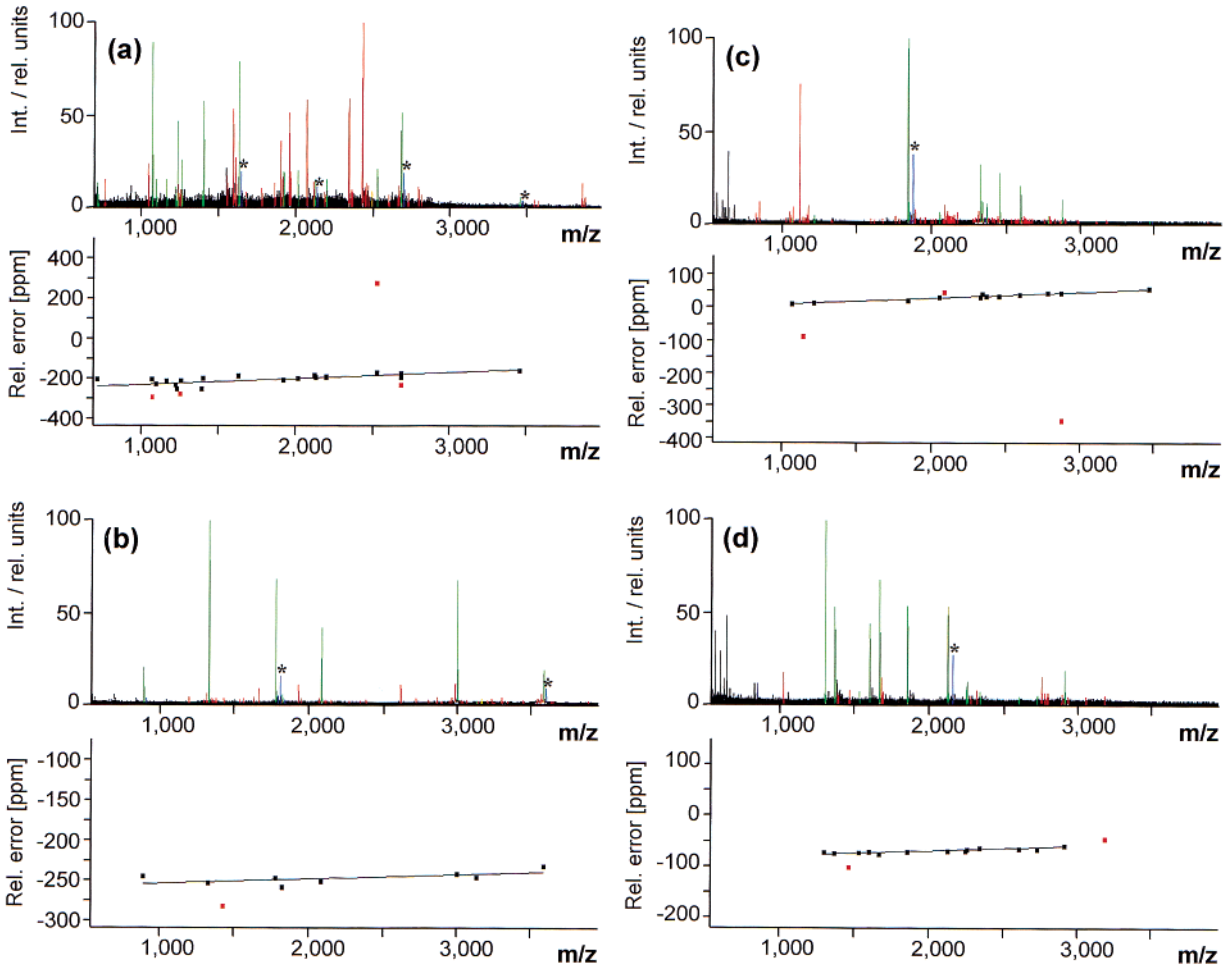

Figure 7. Four examples of different recombinant proteins correctly identified in the NCBI database. (a) Phytochrome B from *A.* thaliana; *Z* score = 99.118, *n* = 19, SD = 14.2 ppm, and SC = 22.3%. (b) Human proteasome subunit α type 1; *Z* score = 99.251, *n* = 8, SD = 5.5 ppm, and SC = 57.4%. (c) Human mitotic centromere-associated kinesin; *Z* score = 99.740, *n* = 12, SD = 2.5 ppm, and SC = 33.4%. (d) Human bleomycin hydrolase; *Z* = 99.868, *n* = 13, SD = 2.2 ppm, and SC = 49.2%. The matched signals are green in the corresponding MALDI-TOF-MS tryptic peptide mass spectra, reproduced in (a−d). Red signals could not be assigned, and blue signals marked by a star indicate oxidized methionine or tryptophan residues. Assigned outliers that matched the correct protein sequence within ±500 ppm, allowed in the initial database search, are marked red in the relative error plots below the spectra.

cases the next following, nonrelated sequence in the ranking list was different and yielded a higher score. An example is shown in Figure 8. In this case, putative triosephosphate isomerase from *A. thaliana* was correctly identified with both peak lists, one containing 78 and the other 161 masses. The longer list yielded 5 hits more for the correct sequence (19 versus 14) and 8 more for the highest ranking, nonrelated sequence (12 versus 6). Although five additional masses were assigned to tryptic cleavage products of the correct sequence for the long peak list, increasing the SC from 60.3% to 67.9%, the *Z* score changed only little (99.63 versus 99.72). The reason for this is that for the calculation of *Z*, the gain in number of hits and SC was mostly neutralized by a decline of the SD from 8.5 ppm for the short list to 12.6 ppm for the long list. Inspection of the additionally assigned peaks revealed that in all cases the signal-to-noise ratio was less than 2.0 and in two cases the peak shape was considerably distorted. Inclusion of these peaks degraded the correlation of the submitted to the matched masses. This was recognized and taken into account by the scoring algorithm, with the consequence that more hits were required for identification.

False Positive Results. If, in the above experiment, the search was extended to all protein sequences contained in the NCBI database, several nonrelated false sequences received a score above 98 and close to 99 (data not shown). In no case, however, was a nonrelated false sequence ranked higher than the correct one or a homologue to it. In the following, the term *false positive* refers to protein sequences that are not homologues of the correct sequence. With respect to the number of submitted masses, we cannot specify a cutoff above which false positive results have to be taken into account, because many if not most or all of the submitted "false" masses are not random. Prominent reasons are cleavage products of a second protein, which was not or incompletely separated, nonspecific cleavage products, fragment ions, and posttranslational or secondary modifications. We conclude that our strategy is tolerant toward "false" masses, but if the database is very large and no efficient restrictions can be put on the search, this tolerance should not be stressed with very long peak lists. In the other experiments reported here, a maximum 100 signals were assigned in each spectrum.

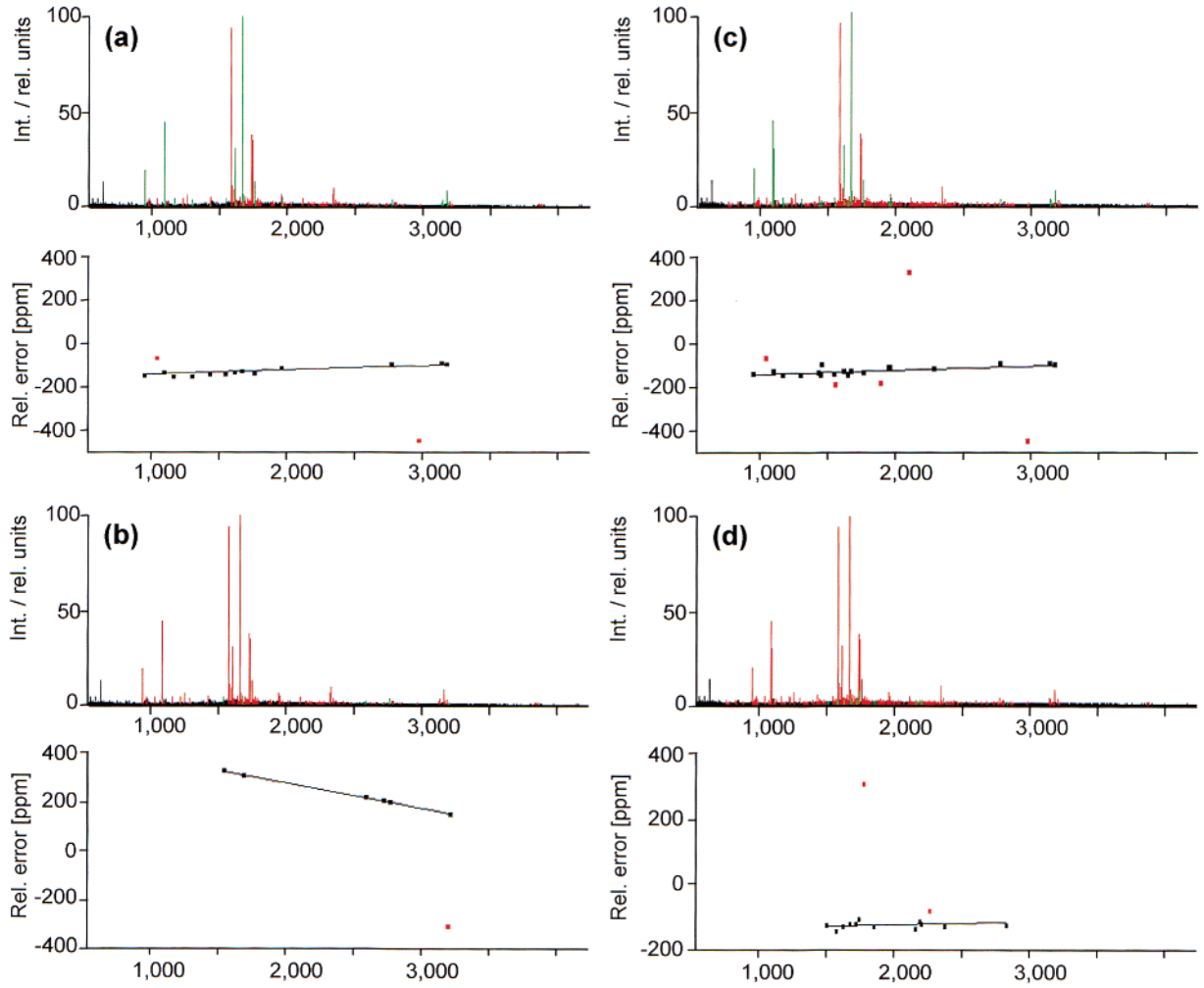

Figure 8. Influence of the number of submitted masses. A tryptic digest of putative triosephosphate isomerase from *A. thaliana* was analyzed by MALDI-TOF-MS, and the recorded spectrum was processed two times. In one run, flight times were assigned to 78 signals, green or red in the spectrum reproduced in (a−d), based on which the correct protein sequence was identified in the NCBI database. (a) The recorded spectrum with the matching signals are green. Below, relative error plot. Red squares, outliers that matched the correct protein in the initial database search within ±500 ppm. (b) As (a) but for the second highest ranking sequence, not related to triosephosphate isomerase. In a second run, 161 signals were assigned and used for identification. (c, d) As (a, b) but for the results of the second spectrum analysis.

Figure 9 summarizes the results obtained for a tryptic digest of the human heat shock cognate 71-kDa protein, recombinantly expressed in *E. coli*. Searching all mammalian protein sequences contained in the NCBI database, the correct sequence (NCBI, 5729877) was identified with a *Z* score of 99.833 ($n = 22$, SD = 9.5 ppm, SC = 58.7%). The second highest ranking, nonrelated candidate was the sequence of a large protein, human myosin heavy chain (cardiac muscle, β isoform, MW = 223 600). For this sequence, the *Z* score was calculated to 98.829 ($n = 31$, SD = 54.7 ppm, SC = 24.3%), a value that suggests this sequence to be a possible candidate, although it is not related to the correct sequence. This example for a false large protein that yields a *Z* score close to 99 belongs to the most extreme we have observed in our studies. The reason for the high score is that a large number of false hits ($n = 31$) remained after removal of six outliers. The corresponding error plot reproduced in Figure 9c shows the reason; instead of following a linear trend line, the data points form a broad cloud, for which the SD to the corrected linear fit was calculated to 54.7 ppm. For the calculation of the *Z* score, this high value and a low SC (24.3%) mostly compensated the impact of the large number of matching peptide masses ($n = 31$), which otherwise (ranking according to n) would have yielded a false identification.

In our studies, we have never observed a standard deviation of >25 ppm for the correct sequence, even if the signal-to-noise-ratio of the assigned signals was very low. An example is documented in Figure 5. Based on this experience, to exclude false-positive large proteins from the ranking list, in addition to increasing the *F* factor (eq 2), the program MSA allows one to limit the maximum SD accepted for the calculation of *Z*. The default value is 30 ppm, which if applied would have excluded human myosin heavy chain as a false candidate from above ranking list.

We have explored the possibility of identifying more than one protein in one sample in the NCBI database. For this purpose, digests of different nonrelated recombinant proteins were mixed and analyzed simultaneously. As expected, the results were not consistent. In many but not all cases, both proteins were suggested a likely candidate (data not shown). It is clear that the risk for a false positive result increases with the number of proteins

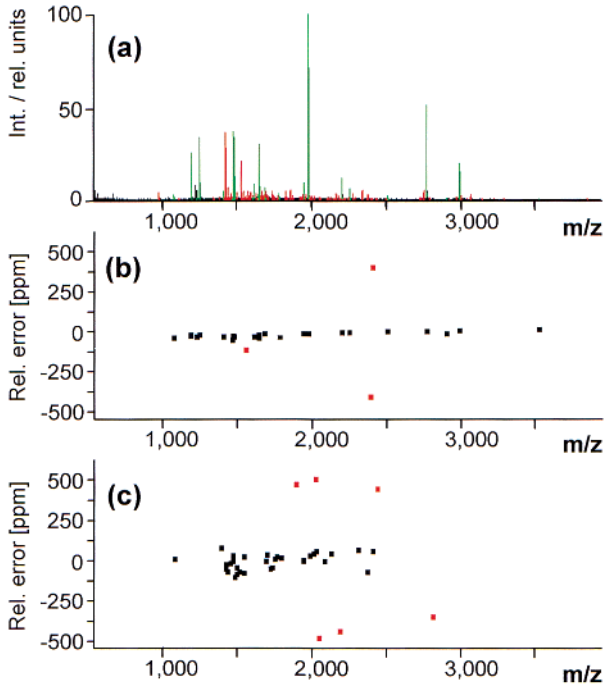

Figure 9. Discrimination of large, false-positive protein sequences. (a) MALDI-TOF mass spectrum of a tryptic digest of recombinant human heat shock cognate 71-kDa protein, used to identify the protein in the NCBI database. (b) Relative error plot for the correct protein sequence, which was identified with a *Z* score of 99.833 ($n = 22$, SD = 9.5 ppm, SC = 58.7%) Red squares, outliers assigned by the program MSA, which matched an expected peptide mass within ±500 ppm. (c) Relative error plot for human myosin heavy chain, the in the generated ranking list next following, nonrelated candidate. Although 31 of the determined masses matched this protein with a molecular weight of 223 600, was the calculated *Z* score insufficient for identification.

contained in the sample. Therefore, we do not recommend acceptance of identification of several nonrelated proteins without verification, e.g., by an independent database search based on MS/MS data.

False Negative Results. The minimum requirements for a likely identification are $n \geq 5$, SD < 30, and $Z \geq 99$ (default settings). As outlined in Figure 6, these conditions are very stringent. For instance, if the SD is 10 ppm and the SC 20%, a minimum 17 hits are necessary to raise the score above 99. If the SC is 50% instead, this demand drops to 11 hits, and if instead of a higher SC, the SD drops to 5 ppm, a minimum 12 hits are required. The advantage of such stringent conditions is that false positive results are rare (not observed in our studies). The price for this is false negative results. We consider the latter category less dramatic than the first and, therefore, use very stringent conditions. If necessary, however, the stringency can be relaxed (or further increased) by modifying the factor *F* in the scoring function. An alternative to the conditions used in this study provides the following settings: $n \geq 6$, SD < 30, $Z \geq 99$, and $F = 0.8$. As a consequence of reducing *F* from 1.0 to 0.8, less hits and a lower sequence coverage or higher values for SD are accepted for identification. This is counteracted by increasing the minimum number of hits from 5 to 6, reducing the number of candidate sequences and the risk for false positive results based on very low values for SD.

ACKNOWLEDGMENT

We thank Prof. J. Klose, Institute for Human Genetics, Virchow-Klinikum, Humboldt-University, Berlin, Germany, for provision of the native protein material used. Dr. Kersten, Max-Planck-Institute for Molecular Genetics, Berlin, we thank for provision of purified recombinant *A. thaliana* proteins and K. D. Kloeppel for assistance with the mass spectrometric instrumentation. This work was funded by the German Ministry for Education and Research and the Max-Planck-Society and was performed in partial fulfillment of the doctoral thesis of V.E. and P.G..

Received for review November 20, 2001. Accepted February 12, 2002.

AC011204G