Algorithm for Accurate Similarity Measurements of Peptide Mass Fingerprints and Its Application

Flavio Monigatti and Peter Berndt
F. Hoffman-La Roche Ltd., Roche Center for Medical Genomics, Basel, Switzerland

We present a simple algorithm which allows accurate estimates of the similarity between peptide fingerprint mass spectra from matrix assisted laser desorption/ionization (MALDI) spectrometers. The algorithm, which is a combination of mass correlation and intensity rank correlation, was used to cluster similar spectra and to generate consensus spectra from a data store of more than 100,000 spectra. The resulting first spectra library of 1248 unambiguously identified different protein digests was used to search for missed cleavage patterns that have not been reported so far and to shed light on some peptide ionization characteristics. The findings of this study could be directly implemented in peptide mass fingerprint search algorithms to decrease the false positive error rate to $<0.25\%$. Furthermore, the results contribute to the understanding of the peptide ionization process in MALDI experiments. (J Am Soc Mass Spectrom 2005, 16, 13–21) © 2004 American Society for Mass Spectrometry

The combination of high throughput protein separation with automated mass spectrometric protein identification has emerged as one of the most useful working techniques in proteomics. The rapidly growing amount of mass spectrometry data requires more reliable and highly accurate methods that help in managing the data. Novel methods for analyzing mass spectrometric data are required for a number of biochemical research methods, including genomics (DNA sequencing), proteomics (protein identification), metabolomics (tracing metabolites), and imaging.

MALDI-MS peptide mass fingerprinting of isolated proteins with peptides derived from endoproteolytic digests has been widely established as the technique of choice for high throughput protein identification [1–3]. After annotation and calibration of mass spectrometric data, the observed peptide masses are compared with molecular weights derived from an in-silico digest of a sequence database. This mapping has necessarily probabilistic character, with the likelihood and the reliability of the match dependent on the mass accuracy of the data, the validity of the used sequence databases, and the employed cleavage and modification patterns. Most currently available search algorithms do not take into account information about peak intensities, individual peak mass errors, favored missed cleavages, and composition of amino acids within the matched peptide sequences [4–6]. By including knowledge of the latter parameters, peptide mass fingerprint (PMF) search algorithms can become more sensitive and/or more accurate.

As an example for the need to include additional knowledge about the peptides we consider the problem of including missed cleavages. Peptides with missed cleavages are experimentally observed peptides that contain a potential protease cleavage site within the sequence. For example, it is well known that trypsin does not cleave when lysine is followed by a proline [7]. While it is desirable to consider these peptides in peptide mass fingerprint searches, it is, on the other hand, advantageous to include only frequently missed cleavage pattern into the searched digest database. The inclusion of superfluous peptides into the search database increases the background for the identification, leading to a diminished sensitivity for the same accuracy of matches.

Unfortunately, the big potential of data-mining large databases of available mass spectrometric data has so far not been realized. In one of the few studies available, Thiede et al. [7] described some of the missed tryptic cleavages that are observed with a higher frequency than expected. Krause et al. [8] presented the dominance of arginine-containing peptides in MALDI-MS peptide fingerprints. The effects of secondary structures on signal intensities have been investigated by Wenschuh et al. [9]. The results of these studies are all derived from a few experimentally validated spectra.

In this article, we analyzed a large body (more than 100,000 fingerprint mass spectra) of experimental data acquired in our group during the last three to four years. First, we propose a method to group together similar spectra and to extract information from all the spectra in order to generate a representative (consen-

Published online November 18, 2004
Address reprint requests to Dr. P. Berndt, F. Hoffman-La Roche Ltd., RCMG Bldg. 093/4.56, 4070 Basel, Switzerland. E-mail: Peter.Berndt@Roche.com

1044-0305/05/$30.00
doi:10.1016/j.jasms.2004.09.013

Received June 29, 2004
Revised September 8, 2004
Accepted September 12, 2004

sus) spectrum for this group. Instead of searching every spectrum against peptide databases, more reliable identifications are rapidly obtained using the database of consensus spectra derived from our clustering procedure. Identification provides knowledge of two different categories of peptides: peptides of which the corresponding mass has been found, and peptides of which the mass is not observed in any experimental spectrum. These two groups of peptides, the matched and the unmatched, are stored in a database along with the corresponding proteins and consensus spectra. This follows the rather common concept of creating reference libraries and compare experimental spectra with the reference mass spectra, e.g., the small molecule mass spectra database that is maintained by the National Institute of Standards and Technology (NIST) [10, 11]. In a recent paper, Beer et al. [12] presented a study on how to create such a reference library for MS/MS peptides. However, to our knowledge, the present paper is the first attempt to describe the creation of such a library for MALDI peptide mass fingerprint data.

We used a reference library of more than 1200 unique consensus spectra extracted from the clustering of more than 100,000 MALDI fingerprints to compare the matching peptides to the unmatched peptides, testing different properties like missed cleavages, hydrophobicity, amino acid distributions, and the tendency to form secondary structures. This allowed for tests of various hypotheses that were put forward in recent literature e.g., about influence of peptide hydrophobicity on signal intensity [13]. To elucidate how the amino acids are distributed at every position in the peptide sequence, we calculated the relative entropy [14, 15] of each amino acid at every position and the potential of the peptides to form secondary structures as described by Williams et al. [16] and Wilmot and Thornton [17].

The results of our analysis have been directly applied in our in-house peptide mass fingerprinting algorithm and resulted in substantial increases in identification accuracy.

Methodology

Mass Spectra

Spectra were taken from a database of mass spectra of tryptic digests of proteins picked from 2D gels. All protein spots were automatically excised, digested using established protocols [18], and analyzed on Bruker Ultraflex instruments (Bruker Daltonics, Bremen, Germany), using ACTH and bradykinin as internal mass standards. All spectra were acquired using reflector mode in a mass range between 850 and 4200 Da. Typically, an automated data acquisition method was used that would accept a spectrum if after 200 shots a minimum peak height and resolution were obtained. As explained below, monoisotopic peptide masses were automatically detected from the mass spectra, filtered for known keratin, trypsin, and matrix fragments, and compared to theoretical masses of peptides derived from an in-silico tryptic digest of all proteins from protein sequence databases (e.g., SwissProt, or NCBI human, mouse, or rat genome draft, as appropriate).

Peak Annotation for MALDI Mass Spectra

Mass spectrometric data was filtered two times using a low-pass median parametric spline filter in order to determine the instrument baseline. The smoothed residual mean standard deviation from the baseline is used as an estimate of the instrument noise level in the data. After baseline correction and rescaling of the data in level-over-noise coordinates, the data point with the largest deviation from the baseline is used to seed a non-linear (Levenberg-Marquardt) data fitting procedure to detect possible peptide peaks. Specifically, the fit procedure attempts to produce the best fitting average theoretical peptide isotope distribution parameterized by peak height, resolution, and monoisotopic mass. The convergence to a significant fit is determined tracking σ values [21]. Convergence is reached if σ does not change more than 0.1 for five successive iterations. After a successful convergence, an estimate for the errors of the determined parameters is produced using a bootstrap procedure utilizing sixteen repeats with a random exchange of 1/3 of the data points. The resulting fit is subtracted from the data, the noise level in the vicinity of the fit is adjusted to the sum of the extrapolated noise level and the deviation from the peak fit, and the process is iterated to find the next peak as long as a candidate peak more than five times over level of noise can be found. The process is stopped when 50 data peaks have been found. The zero and first order of the time-of flight to mass conversion are corrected using linear extrapolation from detected internal standard peaks, and confidence intervals for the monoisotopic mass values are estimated from the mass accuracies of the peaks and standards. Commercially available algorithms, such as Bruker Daltonics SNAP algorithm, can be substituted for this algorithm.

Probabilistic Matching of Spectra Peaks to In-Silico Protein Digests

Peak mass lists for mass spectra are directly compared with theoretical digests for whole protein sequence databases. For each theoretical digest, $[1 - \Pi(1 - N\ P\{p_i\})]^{cMatches}$ is calculated, where N is the number of peptides in the theoretical digest, $P(pi)$ is the number of peptides that match the confidence interval for the monoisotopic mass of the peak divided by the count of all peptides in the sequence database, and *cMatches* is the number of matches between digest and mass spectrum. It can be shown that this value is proportional to the probability of obtaining a false positive match between digest and

spectrum. Probability values are further filtered for high significance of the spectra peaks that produce the matches. After a first round of identifications, deviations of the identifications for mass spectra acquired under identical conditions are used to correct the second and third order terms of the time-of-flight to mass conversion. The resulting mass values have mostly absolute deviations less than 10 ppm. These mass values are then used for a final round of matching, where all matches having a P_{mism} less than $0.01/\mathrm{N}_{\mathrm{Proteins}}$ are accepted.

The Datasets

We generated datasets totalling 100,000 spectra. The samples were derived from four different organisms:

- ~35,000 spectra from human HEK293 cell line
- ~15,000 spectra from *B. subtilis*
- ~14,000 spectra from *Paracoccus z.*
- ~18,000 spectra from Rat *insulinoma* cell line INS1
- ~18,000 spectra from human blood plasma

The cell line samples use were from conventional, untreated, log-phase cultures. We generated datasets of more than 100,000 spectra. All the spectra have been treated as described above.

Spectra Similarity

All spectra which we analyzed have been calibrated and their peaks have been annotated. After theses procedures, we assumed that all deviations from true values in the data were due to non-systematic deviations, and treated a peak as sampling from a normally distributed value. Therefore, a spectrum can be treated as the sum of Gaussians. The discrete correlation theorem states that the discrete correlation of two real functions g and h is one member of the discrete Fourier transform pair $\mathrm{Corr(g,h)_j} \leftrightarrows \mathrm{G_k H_k}^*$ where $\mathrm{G_k}$ and $\mathrm{H_k}$ are discrete Fourier transforms of $\mathrm{g_j}$ and $\mathrm{h_j}$, and the asterix denotes complex conjugation. Therefore, the correlation between two spectra is the inverse Fourier transform of the product of the Fourier transform of the first spectrum with the complex conjugate of the Fourier transform of the second spectrum. For sums of Gaussians, this can be calculated analytically to yield the mass correlation function mc(x) (eq 1).

$$mc(x) = \sum_{i=1}^{i=N_1} \sum_{j=2}^{j=N_2} \frac{S_i S_j}{\sqrt{S_i^2 + S_j^2}} \; e^{-\frac{(m_{\mathrm{oi}} - m_{\mathrm{oj}} - x)^2}{2(s_i^2 + s_j^2)}} \tag{1}$$

In this formulation, each peak (i) from one spectrum is compared with all the peaks (j) of the other spectrum. Spectrum one contains N_1 peaks and spectrum two N_2. The exponent becomes zero when the monoisotopic masses (m_0) of peak i and j are exactly the same at a lag $x = 0$. s_i and s_j denote the standard deviations of the peaks. While the eq 1 was derived for the general case, for calibrated mass spectra per definition the lag time is zero. The above formalism consciously treated spectra assuming that all the peaks have the same intensity.

In order to include intensity information in spectra comparisons, builders of spectra libraries usually employ the angle between spectra normalized to unit length [19, 20]. In the case of MALDI mass spectrometric data, however, reproducing exact intensities over a whole spectrum is very difficult, as slight changes in laser power, acquisition parameters, or crystallization parameters can alter the spectrum. In this article, we propose to use a more robust procedure to include intensity information into spectra comparisons: rank correlation.

In order to compare two spectra, we choose all peaks in a spectrum that have overlapping peaks in the comparison spectra and vice versa. If we assign the highest overlapping peak in a spectrum the rank 1, the second highest rank 2, and so on, then the Spearman rank correlation coefficient of two spectra is defined as:

$$r_c = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}} \tag{2}$$

R_i is the rank of peak i in spectrum one and S_i is the rank of peak i in spectrum two. $\bar{\mathrm{R}}$ and $\bar{\mathrm{S}}$ are the midranks.

In our procedure, if two peaks have intensities that are within 10% of each other, we assign the same rank (tie). The actual calculation of the rank correlation coefficient is conveniently done in terms of the sum-squared difference in ranks as described in [21].

Mass and rank correlation are supposed to be orthogonal (as they use different information from the spectrum). Thus, they should be equally weighted (see the Results section), and we can combine the mass correlation and the rank correlation by calculating their geometric mean.

Clustering Algorithm

We grouped similar spectra together by a clustering method based on the similarity measure described above. Two different datasets can be clustered by generating a trigonal matrix containing all the correlation values obtained by pairwise comparison of spectra. Before clustering, we use an additional simple noise filter removing all peaks in the spectra set that occur in more than 40% of the spectra.

Using the correlation values, we construct a graph by connecting spectra with the highest available correlation value that will not connect to a spectrum that has already been linked to spectra in the current component of the graph in order to prevent it to become circular.

Thus, our procedure constructs a forest of minimal spanning trees with primitive natural chain breaks. This kind of data structure allows walking along the graph, from the shortest distance between two nodes to the

farthest node, visiting each node at least once. This procedure also reduces the dimensionality of the problem from two dimensions (the trigonal matrix) to one dimension, allowing for highly efficient processing of large amounts of data.

As only the highest scorers get the chance to be connected, we will obtain clusters that are maximal consistent. Nevertheless, we still encountered problems in the grouping process due to spectra that contained peaks from two different proteins. Protein mixture spectra act as bridging edges between two independent clusters. To prevent such inconsistent cluster formation due to protein mixtures, we have introduced a variable threshold parameter. The threshold parameter setting is varied in an iterative process where the cluster consistency is evaluated each time. Iteration is stopped when all resulting clusters are uniform.

Database of Consensus Spectra

Clusters obtained by the procedure described above can also be seen as a list of spectra that contain similar peaks. An obvious exertion of this list would be the generation of a consensus spectrum. Such a spectrum contains the most abundant peaks that overlap in spectra assembled in the cluster. We typically use the first 50 most frequent peaks that occur in more than 20% of the spectra. From a consensus spectrum, average peak molecular masses and their standard deviations can be calculated. In addition, a rank order estimate can be derived. When a consensus spectrum is identified by peptide mass fingerprinting, a sequence can be assigned to the peaks by calculating a theoretical tryptic digest and comparing masses of the digest with masses of the consensus spectrum. Doing that for all consensus spectra results in two groups of peptides from the theoretical digests, peptides of which masses correspond to measured peak masses and peptides of which masses do not correspond to any of the peaks. Matched and unmatched peptides are stored in a database along with the consensus spectra and the underlying protein identifications.

Peptide analysis: We compared the matching peptides with the unmatched ones, testing different properties like missed cleavages, hydrophobicity, hydrophobicity gradient, amino acid distributions, and the tendency to form secondary structures. The hydrophobicity of a peptide was calculated by summing the per amino acid hydrophobicity score according to Kyte and Doolittle [22]. To obtain information on the amino acid distribution at every position in the peptide sequence we calculated the relative entropy of each amino acid at every position. Relative entropy can be seen as the information content in "bits", when the distribution of the experimentally determined amino acid frequency at a certain position is compared to its background distribution, e.g., the amino acid composition of the SWISS-PROT database. It can be used to measure the degree of conservation at a site in a peptide sequence alignment.

In a last test we analyzed the potential of the peptides to form secondary structures. Three categories are differentiated: α-helix forming peptides, β-sheet forming peptides and peptides that do not tend to form one of the preceding conformations. The rules of how to classify the peptides into these categories where defined by Williams et al. and Wilmot and Thornton [16, 17].

Results

A Novel, Accurate Similarity Measure for MALDI PMF

Similarity measures used in construction of mass spectrometric libraries either ignore the intensity component of the mass spectrum or use different normalization schemes to normalize the ion current. We have evaluated these measures by pair wise comparisons of a set of spectra of which the underlying protein was well known. The set contained 558 spectra. As shown in Figures 1 and 2, the traditional methods show significant overlap between the two groups of spectra.

We have, therefore, introduced a new similarity measure for MALDI PMF: we combine the correlation of the molecular weights with the Spearman rank correlation of the peak heights. As shown in Figure 3, this measure achieves nearly complete separation of real identities from random matches. We have tried various weights and averaging procedures for the combination of the two correlation measures, and we have established that the best method to combine mass and rank correlation is to calculate their geometric mean.

Clustering Method

Using the similarity measure established above, we have clustered more than 100,000 spectra using a simple closest neighbor graph constructing algorithm. Around 80% of the spectra could be assembled in clusters by the

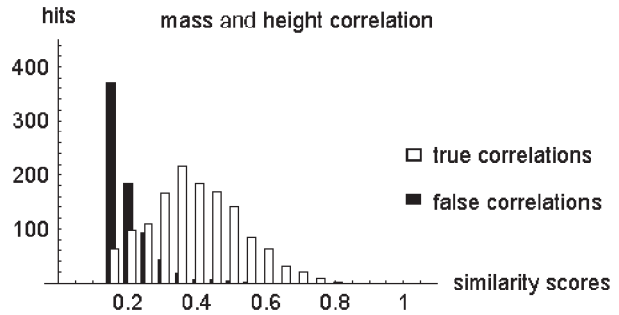

Figure 1. Pair wise spectra comparisons using mass spectra with intensities normalized to unit vector [19, 20] show a significant overlap of scores for true and false correlations for the same spectra sample as in Figure 1. The difference between the median false correlation bin and its true correlation counterpart is only 0.2. While calculating the rotational angle between vectors in mass space is the usual approach to construct mass spectra libraries, it is clearly not a satisfactory method for separating MALDI PMF spectra.

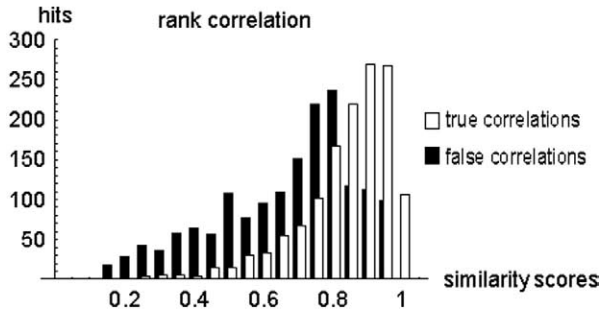

Figure 2. Pair wise spectra comparison of the spectra set used in Figure 1 was scored using the Spearman rank correlation coefficient alone. The separating power of this method is even lower than with the rotational angle method (Figure 2).

clustering method. The clustering resulted in a set of 5167 clusters from which a consensus spectrum could be extracted. 1713 of these consensus spectra in turn led to 1827 protein identifications, of which 1248 were unique. 933 identifications occurred more than once. As our means of generating consensus spectra tries to include the most abundant peaks in the spectra from the clusters, the number of peaks does not reflect the number of theoretical peptides. Considering the completeness of sequence databases employed for searching, we would estimate that ~60% of the consensus spectra should yield identification. Our identification rate is somewhat lower, a fact that is explained by the consideration, that only for large clusters (where the number of assembled spectra is greater than ~10) a significant improvement of the peak parameters can be expected. The peak–peptide mappings resulted in a total of 11235 unique peptide sequences, with an average of 9 peptides per consensus spectrum. All collected data were stored in a database. These clusters have been mapped to the original 2D PAGE gels. Additionally, cluster accuracy was evaluated on the basis of its graphical representation. An energy minimized drawing of an α-enolase cluster using a 2D spring embedding algorithm [23] is shown in Figure 4.

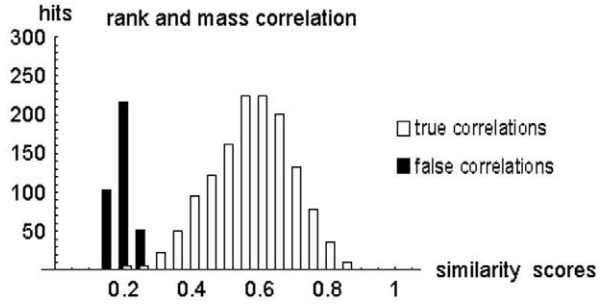

Figure 3. Combining mass and rank correlation of MALDI mass spectra allows complete discrimination between truly related spectra and unrelated ones in this sample of 558 spectra. The difference between the median false correlation bin and its true correlation counterpart is around 0.4.

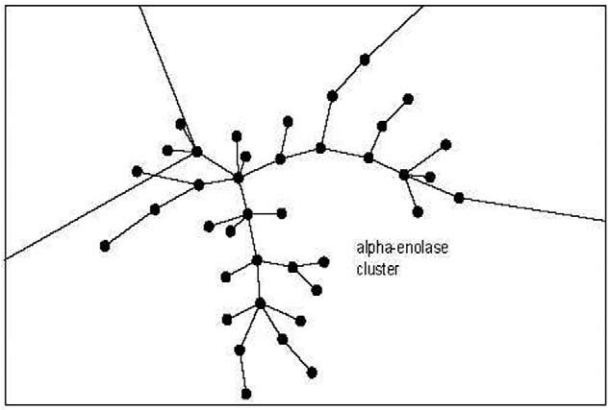

Figure 4. Graph of a single cluster of 38 spectra isolated from a set of ca. 1500 MALDI PMFs. The core region consists of spectra that have been identified as αenolase. Smaller distances correspond to stronger similarities. Note that the average distances within the cluster are similar and very distinct from the outbound connection distances. A iterative procedure has been implemented that employs this feature to separate self-consisting clusters.

Missed Cleavage Patterns

We used previously established databases of consensus spectra from *Bacillus subtilis* and human samples. With the peak lists of the consensus spectra we carried out database searches to identify underlying proteins without considering missed cleavages or posttranslational modifications. If a spectrum was identified, it was compared to the theoretical digest of the identified protein again, allowing, this time, two missed cleavages within the sequences.

In total, we have analyzed 3251 theoretical peptide sequences from *Bacillus subtilis* and 6859 theoretical sequences from human proteins. These 10,110 sequences contained 5760 missed cleavage sites which were subjected to an analysis of the two cleavage site flanking amino acids. Results are shown in Table 1.

Distribution of R- and K-Ending Peptides Within the First Ranks

The database of consensus spectra includes the rank order of each match. We queried the database for the occurrence of lysine- or arginine-ending peptides within the first few ranks of the consensus spectra. The result of that investigation is shown in Figure 5.

As it is obvious from the data, arginine-ending peptides are much more frequent among the top-ranking

Table 1. Missed cleavage patterns accounting for >90% of the detected miscleavages in our sample of tryptic protein digests

Pattern 1	[WYF][RK][^RK] or [^RK][RK][WYF]
Pattern 2	[DE][RK][^RK] or [^RK] [RK][DE]
Pattern 3	[^RK][RK][RKH]
Pattern 4	[RK][P]

The letters (amino acids) in brackets mean that at least one of them has to occur. A pattern like WKA would fall into the first category.

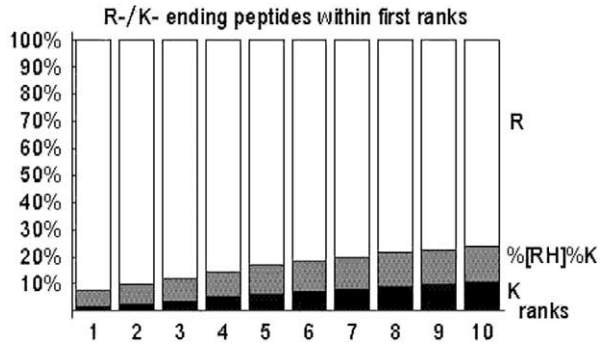

Figure 5. Distribution of R-/K- terminated peptides within the *n* rank most intense MALDI signals. K-ending peptides without other basic residues are almost absent. K-ending peptides that either contain a histidine anywhere in the sequence or a missed cleavage site consisting of an arginine occur more often than lysine-ending peptides that do not have the mentioned sequence characteristics.

peptides then their statistical expectancy, with the frequency of observing a K-ending peptide without another basic residue among the first 10 ranks being less than 8%. This holds for multiple lysine residues in the sequence. As shown in Figure 6, the occurrence of multiple lysine residues in matched peptides without arginines or histidines is distributed approximately like a binomial with a frequency of 6.5% (instead of 27.8%, as expected by amino acid frequencies).

Kyte and Doolittle Hydrophobicity Plot

We have evaluated the hydrophobicity and the gradient of hydrophobicity for each peptide in the database. This was done using the per amino acid hydrophobicity score determined by Kyte and Doolittle [22]. No difference between the matched and unmatched peptides could be observed. Additionally, we calculated the hydrophobicity and the hydrophobicity gradient in dependency on peptide rank for the ten top-ranked peptides. We could not observe positional dependencies (amphiphilic character) for the hydrophobicity of the amino acid side chains (data not shown).

Amino Acid Distributions as Indicators of Peptide Ionization Characteristics

For the calculation of the relative entropy per position and amino acid, we took into account the last 10 amino acids on the C-terminal site of the peptides. We assumed that ten positions before the proton accepting Arginine or Lysine are sufficient to reveal possible differences in amino acid distributions comparing matched peptides with unmatched peptides. 8500 peptides containing at least 10 amino acid residues have been aligned starting at the C-terminus. Entropies were calculated by taking the frequencies of amino acids at every position in the alignment. The result of this analysis is shown in Figures 7 and 8. We observed significant differences in residue distribution patterns between matched and unmatched peptides. Peptides detected in consensus MALDI MS spectra contain a marked excess of small amino acids (alanine, glycine, valine), while the occurrence of acidic amino acid residues in the immediate neighborhood of the C-terminal basic amino acid is reduced.

Secondary Structure Analysis

The results of the secondary structure analysis, scoring the amino acid chains according to the rules defined by Williams et al. and Wilmot and Thornton) [16, 17] are shown in Table 2.

Discussion

In the present paper, we have introduced a new similarity measure to accurately compare mass spectra. The algorithm that we developed takes into account the intensities of the peaks by comparing their relative order. This is done using the rank order correlation coefficient [21]. The similarity measure we use performs better than any similarity measure suggested so far. We have compared our algorithm to the known measure of normalized dot product [12, 19, 20]. In this formulation of the problem only the k highest peaks overlapping between two spectra are taken into account and normalized to the total intensity of overlapping peaks. As shown in Figures 1 and 3, the normalized dot product solution does not work accurately enough when measuring the similarity between MALDI spectra. The combination of the mass correlation and the rank correlation does not show false correlations above a similarity score of 0.3, whereas the mass correlation alone or correlations with normalized ion current yield false correlations with similarity values up to 0.6. On the other hand, our method is the only method tested that did not miss true correlations. While it is obvious that adding intensity information to a spectra comparison is complementing it with orthogonal information content, it is crucial to understand that the absolute height of a

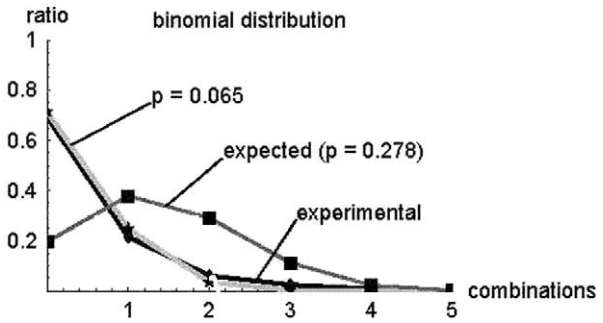

Figure 6. Binomial distribution of K-ending peptides not containing R or H. The experimentally determined values fit to the binomial distribution of values with a probability *P* of 0.065. Theoretical sequence database analysis predicts a 4.3-fold higher abundance of lysine-ending peptides.

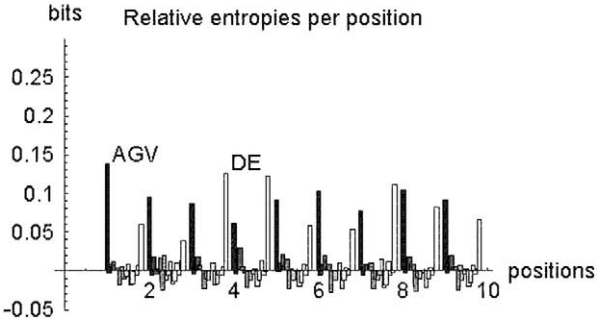

Figure 7. Information content of amino acid residues at the C-terminus of peptides detectable in MALDI PMFs. The entropies for arginine and lysine have been removed because their high information content is caused by the cleavage specificity of trypsin and suppresses the signal of the other positions. Significant differences from the natural amino acid distribution could be detected for small amino acid residues (alanine, glycine, valine) and for acidic residues (glutamic acid, aspartic acid).

MALDI peak is not a robust measure of its quantity. MALDI data is quantitative only in comparison to the other components contained in the same spectrum. Rank order reflects just this fact, and it is robust in the presence of noise or separate components.

We used our new spectra similarity measure to generate similarity graphs by connecting spectra nodes to the nearest node whose edge has not been visited before. The graph is in general disconnected, and contains a forest of trees, each of them representing a cluster of similar MALDI spectra. An iterative cluster consistency checking process guarantees optimal performance. A simple check of the selectivity of our method has been carried out by clustering two datasets from two different organisms (data not shown). Not even a single spectrum from one dataset appears in a cluster of spectra from the other dataset.

The generation of consensus spectra from clusters serves for two purposes: (1) Instead of searching the peptide-database by peptide mass fingerprinting, we

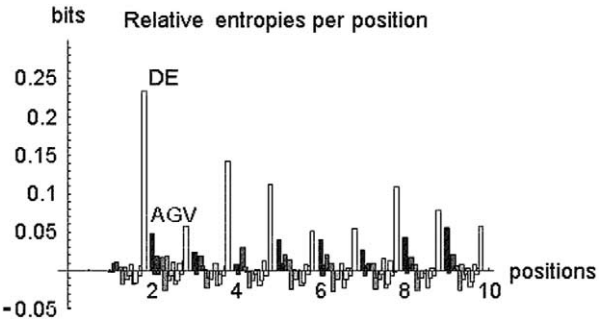

Figure 8. Information content of amino acids residues at the C-terminus of peptides not detected in MALDI PMFs. Significant differences from the natural amino acid distribution could be detected for acidic residues (glutamic acid, aspartic acid) but, contrary to the data presented in Figure 7, not for small amino acid residues (alanine, glycine, valine). As discussed below, this could be interpreted in terms of different structural requirements for ionization in MALDI experiments.

Table 2. Secondary structures

	Matched	Unmatched
H & S	17%	25%
S	32%	22%
C	22%	18%
H	29%	35%

This table shows the tendencies of the two groups to form secondary structures. H stands for helix, S for sheet and C for no secondary structure. The matched peptides tend to form β-sheets whereas the unmatched peptides favor helix conformation.

can compare the spectrum to be identified with the spectra in the consensus spectra database. (2) Consensus spectra can be used to elucidate peptide properties for peaks showing consistently high intensities.

Identification of proteins using a library of consensus spectra has the big advantage of automatically considering only a realistic set of observable peptides. Considering that the sequence coverage in MALDI PMF is typically in the range of 20%, this means that about 80 to 90% of the usually considered peptide masses can be ignored without loss of search accuracy. As easily seen from the probabilistic mismatch estimate, this will lead to several orders of magnitude differences in mismatch probabilities even for few matches and thus increase search sensitivity dramatically. The only drawback of identifying a spectrum by searching the database of consensus spectra is that a consensus spectrum representing the searched protein has to be experimentally observed in previous measurements. Current progress in mass spectrometry instrumentation makes it likely that the task of establishing complete spectra libraries for commonly used laboratory organisms could be feasible in the near future.

As pointed out above, consensus spectra can be used to elucidate constraints on the physico-chemical properties of observed peptides. This serves the purpose of directly reducing the set of peptides taken into consideration for identification purposes.

A dramatic reduction of the size of the peptide database can be achieved by obtaining information on missed cleavages and information on the distribution of K-/R-ending peptides. The four missed cleavage patterns presented in the present study cover 91% of the most often occurring detected missed cleavages. Pattern four is well known because trypsin is unable to cleave when arginine or lysine is followed by a proline. Pattern two and part of pattern three have been described in literature before [7]. Pattern one has not been described in literature before because the three amino acids tryptophan, phenylalanine and tyrosine are low abundant amino acids. Part of Pattern 3, the histidine occurring at the right side of a potential missed cleavage site is a low abundant amino acid too.

The result of the investigation on K-/R-ending peptides allowed us to correctly quantify the well known excess of arginine ending peptides over lysine ending ones [8]; arginine-ending peptides occur up to 20 times

more often within the first few ranks, whereas lysine-ending peptides are almost not seen at these ranks. Peptides having a sequence pattern like '%[RH]%K', which is a peptide that ends with lysine and carries an arginine or a histidine somewhere in the sequence, appear less than peptides having an arginine at the C-terminus, but more than lysine-ending peptides where no arginine or histidine occurs within the sequence. Consistent with the average expectation value of a lysine-terminated peptide without other basic residues, the number of occurrences of these peptides in all consensus spectra follows a binomial distribution with a probability of 0.065 (see Figure 6), whereas the expectation value in theoretical digests is about 0.28. Therefore, false positive protein matches can easily be recognized by their relative excess of lysine-terminated peptides without basic residues.

By including these rules into the PMF search procedure, we lowered its false positive rate substantially. We tested the performance of the algorithm by creating a substitute peptide database derived from a real protein digest database of *Bacillus subtilis* proteins where all the peptide masses are shifted by 3 Daltons. With this database we searched 77,157 experimental spectra and obtained false positive identifications in 122 cases. When searched with the real database, we obtained 49,644 identifications. Thus, our PMF search algorithm is more than 99.75% accurate.

Very little is known about the peptide ionization process in MALDI mass spectrometry. The potential benefit from predicting ionization properties and thus the ranking potential for any given sequence is substantial.

It is widely presumed that the hydrophobicity of peptides influences signal intensities.

However in our sample of over 100,000 spectra, the analysis of hydrophobicities did not show any difference between the matched and the unmatched peptides. Thus, the hydrophobicity of a peptide is most probably not the property that could explain different ionization behavior. On the other hand, analysis of relative entropies of amino acid positions in peptides revealed clear differences between matched and unmatched peptides. Whereas the unmatched peptides mostly carry acidic amino acids before the ion acceptor, the matched peptides prefer small amino acids like gylcine, valine, and alanine at this position. These three amino acids are more prevalent at all ten positions before the C-terminal arginine or lysine. Only at positions four, five, and eight the relative entropies of D and E are higher than the ones for A, G, or V. Such a pattern is not observed in the unmatched peptides, where only D and E residues contain information relative to the background distribution, probably reflecting the prevalence of ion bridges in protein secondary structures.

These empirical results prompted us to draw the hypothesis that the difference in ionization regarding the matched and unmatched peptides could be the result of a tendency to form secondary structures that are different for the two groups. The hypothesis is supported by the fact that the unmatched peptides have a higher α-helix structures index whereas the matched peptides tend to have amino acid compositions more consistent with β-sheet conformations. There are some observations in literature consistent with this point of view [9]. While secondary structure assignment is rather inaccurate, our results still could reflect differential solvation and/or secondary structure formation.

In conclusion, the present work has presented the necessary algorithms and shown that it is possible to distill consistent and reproducible spectra- and peptide information from a database of MALDI MS spectra. We have taken the first steps toward the construction of a spectra library of MALDI PMF spectra. Such a library could revolutionize proteomics data processing in terms of accuracy and speed. Analysis of confirmed matched peptides has lead to the formulation of important constraints on theoretical peptide digest databases used in peptide mass fingerprint matching. Improvement of the experimental base will, without doubt, finally lead to an understanding of the ionization behavior of peptides in MALDI MS.

Acknowledgments

The authors thank all the researchers in the Proteomics group of the Roche Center for Medical Genomics for their enthusiasm in acquiring MALDI MS PMF data. They thank the head of the Proteomics group, Dr. Hanno Langen, for his continuing support and many helpful discussions.

References

1. Lahm, H. W.; Langen, H. Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **2000,** *11*, 2105–2114.
2. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003,** *6928*, 198–207.
3. Zolg, J. W.; Langen, H. How industry is approaching the search for new diagnostic markers and biomarkers. *Mol. Cell. Proteomics* **2004,** *4*, 345–354.
4. Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999,** *14*, 2871–2882.
5. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** *18*, 3551–3567.
6. Zhang, W.; Chait, B. T. ProFound: An expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **2000,** *11*, 2482–2489.
7. Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **2000,** *6*, 496–502.
8. Krause, E.; Wenschuh, H.; Jungblut, P. R. The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal. Chem.* **1999,** *19*, 4160–4165.
9. Wenschuh, H.; Halada, P.; Lamer, S.; Jungblut, P.; Krause, E. The ease of peptide detection by matrix-assisted laser desorp-

tion/ionization mass spectrometry: The effect of secondary structure on signal intensity. *Rapid Commun. Mass Spectrom.* **1998,** *3*, 115–119.
10. Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **1999,** *4*, 287–299.
11. Josephs, J. L.; Sanders, M. Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. *Rapid Commun. Mass Spectrom.* **2004,** *7*, 743–759.
12. Beer, I.; Barnea, E.; Ziv, T.; Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **2004,** *4*, 950–960.
13. Olumee, Z.; Sadeghi, M.; Tang, X. D.; A., V. *Rapid Commun Mass Spectrom*. **1995**, *9*, 744–752.
14. Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological sequence analysis*; Cambridge University Press: Cambridge, 1998, pp 308–309.
15. Shannon, C. E. The mathematical theory of communication, 1963. *MD Comput*. **1997,** *4*, 306–317.
16. Williams, R. W.; Chang, A.; Juretic, D.; Loughran, S. Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta.* **1987,** *2*, 200–204.
17. Wilmot, C. M.; Thornton, J. M. Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **1988,** *1*, 221–232.
18. Fountoulakis, M.; Langen, H. Identification of proteins by matrix-assisted laser desorption ionization-mass spectrometry following in-gel digestion in low-salt, nonvolatile buffer and simplified peptide recovery. *Anal. Biochem*. **1997,** *2*, 153–156.
19. Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 859–866.
20. Alfassi, Z. B. On the normalization of a mass spectrum for comparison of two spectra. *J. Am. Soc. Mass Spectrom.* **2004,** *3*, 385–387.
21. Press, W. H.; Teukolsky, S. A.; Flannery, B. P.; Vetterling, W. T. *Numerical Recipes in C*; Cambridge University Press: Cambridge, 1992, pp 640–642.
22. Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982,** *1*, 105–132.
23. LEDA: Library for efficient data types and algorithms. *Algorithmic Solutions Software GmbH*; 2004.