A New Matching Algorithm for High Resolution Mass Spectra

Michael Edberg Hansen*
Informatics and Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark

Jørn Smedsgaard
BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark

We present a new matching algorithm designed to compare high-resolution spectra. Whereas existing methods are bound to compare fixed intervals of ion masses, the accurate mass spectrum (AMS) distance method presented here is independent of any alignment. Based on the Jeffreys-Matusitas (JM) distance, a difference between observed peaks across pairs of spectra can be calculated, and used to find a unique correspondence between the peaks. The method takes into account that there may be differences in resolution of the spectra. The algorithm is used for indexing in a database containing 80 accurate mass spectra from an analysis of extracts of 80 isolates representing the nine closely related species in the *Penicillium* series *Viridicata*. Using this algorithm we can obtain a retrieval performance of ≈97–98% that is comparable with the best of the existing methods (e.g., the dot-product distance). Furthermore, the presented method is independent of any variable alignment procedures or binning. (J Am Soc Mass Spectrom 2004, 15, 1173–1180) © 2004 American Society for Mass Spectrometry

In recent years analytical chemists have benefited from an impressive development of instruments and especially in their performance. This is particularly true for modern mass spectrometers, by which nearly all types of problems now are addressed. However, a new bottleneck has arisen; the processing and interpretation of the enormous amount of data that can be produced. One way to minimise the tedious work of interpreting huge amounts of spectra is via library search methods. Here, a library of known spectra is coded, often by extracting a subset of data from the complete spectrum. Generally, the spectrum is transformed into a code suitable for computer searching, reducing both storage requirement and search time. Each unknown spectrum is coded as a library entry, and compared with either the complete library or a selected subset to find those spectra entries, which are "best fits" of the unknown according to selected criteria. It is clear that the coding method and "best match" metric chosen will influence the efficiency and quality of the retrieval and identification success. The aim of spectrum evaluation can be either the identification of a compound (assuming a reference spectrum is already available) or the interpretation of spectral data in terms of the unknown chemical structure or comparing spectra of complex mixtures. Identification is most efficiently performed by library search methods based on spectral similarities [1–6]; there are a number of MS databases and powerful software products offered which are routinely used for this purpose. Most of these methods (if not all) are focused on libraries containing mass spectra of pure compounds, but there should be no reason not to store and retrieve complex mass spectra by these library search products [7]. In order to evaluate the propinquity and obtain a "fit" or "match" between spectra, several metrics have been proposed in literature. Fundamentally, these metrics define clusters of spectra and are used to measure the similarities between two patterns from the same feature space. The distance measure (or measures) must be carefully chosen to deal with different data modalities, features extracted, and their scales. In the mass spectroscopic literature, a long list of methods for library construction and search has been proposed since the beginning of the 1970s. Stein and Scott [4] reviewed these and compared and tested five of the most popular algorithms proposed in the literature for library searching—including those implemented in many instrument software packages: Probability based matching (PBM) [8], Hertz similarity index [3], Euclidean distance (L_2-norm), dot-product, and absolute value distance. The best performance was achieved with the dot-product function, which measures the cosine of the angle be-

Published online July 3, 2004

Address reprint requests to Dr. M. E. Hansen, Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark. E-mail: meh@biocentrum.dtu.dk

*Also at Bio-Centrum-DTU, Søltofts Plads, Buildinig 221, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

1044-0305/04/$30.00
doi:10.1016/j.jasms.2004.03.008

Received September 10, 2003
Revised March 14, 2004
Accepted March 16, 2004

tween spectra. The Euclidean distance and the absolute value distance followed this measure. Furthermore, intensity scaling and mass weighting by the square root were shown to be important in the algorithms of the intensity scale being nearly optimal, and the square or cube the best mass weighting power. Several more complex weighting schemes have also been tested, but had little effect on the results [4]. All existing methods for comparing spectra require that variables, in this case the masses, (and the corresponding intensities) of ions in the spectra are aligned to calculate their (dis)similarities. It is most common to calculate the dissimilarity between two patterns using a distance measure defined on this rigid feature space (mass scale). Let us assume that we have two observations (i.e., spectra), i and j, with N and M variables (ion masses), $x_i = \{x_{in}\}$, $n = 1...N$ and $x_j = \{x_{jm}\}$, $m = 1...M$. In the case where $N = M$, we can use one of the most popular distances for continuous features taken from the L_p-norm (denoted $\|\cdot\|_p$)

$$d_p(x_i x_j) = \|x_i - x_j\|_p = \left[\sum_{\forall k} |x_{ik} - x_{jk}|^p\right]^{1/p} \quad (1)$$

in which the metric evaluates the dissimilarity between the two vectors of features element by element. In the case of nominal mass data the alignment criterion $N = M$ is fulfilled by aligning unit mass windows (the truncated masses, often with a small offset of e.g., −0.3 Da). Unfortunately, even if $N = M$ is fulfilled in high-resolution data, these are sampled at a high rate on a continuous mass scale (down to several decimals) allowing several narrow ion peaks per unit mass window. Therefore aligning these data is not straightforward. The most widely used approach is binning, using narrow bins. However, it is not trivial to determine the number of bins (and bin width) and the correct placement of these bins in high-resolution data. As illustrated in Figure 1 we may ask whether the ions A and B in Spectrum 1 belong to the same ion population as ion C in Spectrum 2. If 2 and 1 are sampled at nominal resolution they most likely do, whereas if the spectra is sampled at high resolution they may belong to different ion populations. Therefore, if a binning approach is used these ions should be placed in different bins.

The scope of this paper is to present a general approach for automated comparison and classification of high resolution mass spectra, illustrated by analyzing spectra from direct infusion mass spectrometry of crude fungal extracts, exploiting the full data quality in terms of resolution and mass accuracy both within and across samples for, e.g., metabolomics, chemotaxonomy, sample screening/de-screening, and novelty discovery. Our approach is designed to work on both nominal and high-resolution data, or a mixture of these. However, it must be emphasised that any automated processing requires the use of good laboratory practice, thus the instrument should be tuned properly, the mass scale should be calibrated correctly, and good acquisition procedures should be used (e.g., regarding saturation of the detector system).

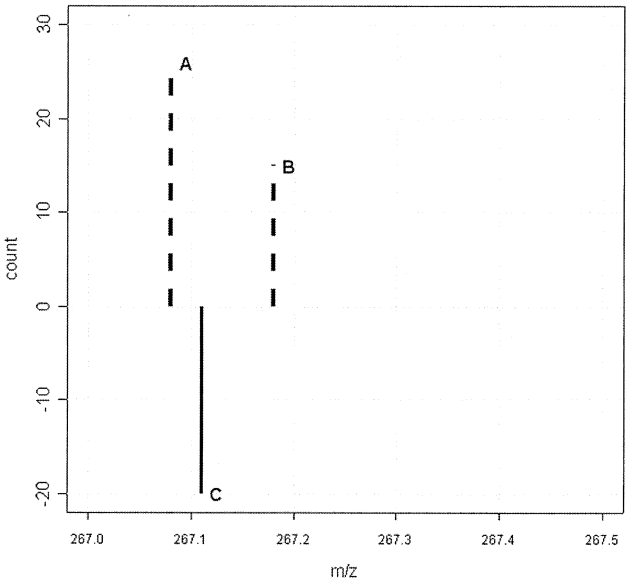

Figure 1. Two spectra containing respectively two (A and B) and one (C) ions (peaks) in the interval of m/z 267.0–267.5. The profile containing one ion has been plotted with negative ion counts for illustrative purpose. By nominal binning A, B, and C will end up in the same bin thus belonging to the same population. Using 0.1 Da wide bins B and C will end up in the same bin, whereas A will be in an other bin even though A and C are closer in mass. Moving the bin structure −0.05 Da will put A and C in the same bin and leave B alone.

Theory

Nomenclature

The following nomenclature will be used in this article: $\boldsymbol{U}$ and $\boldsymbol{R}$ represents two mass spectra, e.g., an *unknown* and a *reference*, of any resolution in normal centroid format. In this general case these could be vectors extracted directly from any instrument, containing all sampled information. But they might as well be continuous data from which the centroid spectra with resolution are extracted.

In order to reduce the computational complexity, each of the spectra are condensed into a set of basic descriptors. These descriptors contain "model" information from the mass spectrum. As an example we choose to extract, e.g., the mass, intensity, and resolution for each of the ion peaks detected in the spectrum.

In the following discussion we use peak for an ion observed as a mass peak in a spectrum. The p'th peak is described by the mass m_p, intensity i_p and peak width w_p (eq 2).

We have used data from direct infusion ESI-MS of complex samples condensed into centroid mass corrected spectra as described in [9]. However it can be any type of mass spectral information. These centroid mass

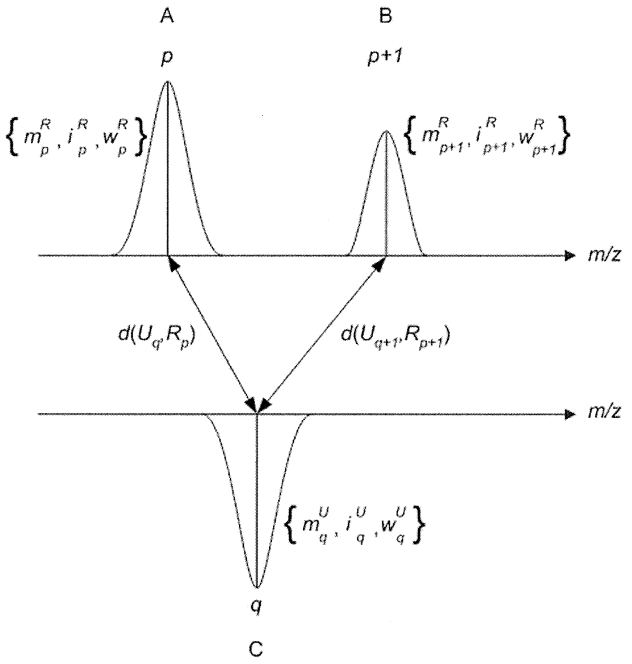

Figure 2. The notation used. From each of the profiles U (lower) and R (upper) information about the peaks is extracted, ϕ^U_q and ϕ^R_p, q = 1, . . . , $|U|$ and p = 1, . . . , $|R|$, in which m is the peak mass, i the intensity, and w is the width of the peak (FWHM).

data are used as input to the algorithm, of which the individual elements in $\boldsymbol{R}$ and $\boldsymbol{U}$ are written as

$$\begin{aligned}\phi^R_P &= \left\{m^R_P, i^R_P, w^R_P\right\} = \{m,i,w\}^R_P \\ \phi^U_q &= \left\{m^U_q, i^U_q, w^U_q\right\} = \{m,i,w\}^U_q\end{aligned} \tag{2}$$

where $p = 1,\ldots, |\boldsymbol{R}|$ is an index running through all peaks in $\boldsymbol{R}$ and $q = 1,\ldots, |\boldsymbol{U}|$ runs through all peaks in $\boldsymbol{U}$. m is the mass (in m/z), i the (log) intensity, and finally w the width of each peak, found as the full width half maximum (FWHM) of the original ion population.

In short, a group of unknown and reference spectra can be written as

$$\begin{aligned}\boldsymbol{\Phi}^R &= \left\{\phi^R_P\right\} \\ \boldsymbol{\Phi}^U &= \left\{\phi^U_q\right\}\end{aligned} \tag{3}$$

A detailed description of the calculation of centroid mass spectra and peak width can be found in [9]. Figure 2 summarizes the notation used, showing a reference spectrum and unknown centroid spectrum with models of the original ion distribution estimated from the parameters in eq 2.

Peak Correspondence

If two spectra, $\boldsymbol{U}$ and $\boldsymbol{R}$, are compared over a mass range, some peaks will be found in one spectrum that are not present in the other, even though their appearances look similar. With reference to Figure 2 the question could be: Does Peak C show more similarities to Peak A or Peak B? In order to answer this question, the first step is to establish a correspondence between $\boldsymbol{\Phi}^U$ and $\boldsymbol{\Phi}^R$.

We assume that the centroid mass spectra are sampled according to good laboratory practice, instruments properly calibrated and mass scale drift minimized e.g., by the use of internal mass reference, see [9].

Comparing the peaks in $\boldsymbol{U}$ with those in $\boldsymbol{R}$, the first assumption is that IF there is a correspondence between two peaks, ϕ^U_q and ϕ^R_p, this will be the only one! Thus, the peaks can, if possible, be combined pair wise. Alternatively, if there is NO correspondence between ϕ^U_q and ϕ^R_p, then ϕ^U_q is absent in $\boldsymbol{R}$.

Therefore, the correspondence between $\boldsymbol{U}$ and $\boldsymbol{R}$ is found by

$$min \sum_{\substack{q=1\ldots|\mathbf{U}| \\ p=1\ldots|\mathbf{R}|}} d_{qp} \tag{4}$$

under the constraints of

1. having p and q assigned only once;
2. peaks cannot switch place;
3. that $d_{qp} \leq d_{min}$.

Here d_{min} is defined as the "distance" at which peaks are regarded as not being the same.

The peak-to-peak correspondence is defined as a distance function, $d_{qp} = d(\phi^U_q, \phi^R_p)$. This function can be chosen in many ways, and should reflect the "distance" between peaks, typically so that $d_{qp} \in [0, 1]$, where 0 means identical peaks (zero distance) and 1 that they are completely different. However, this interval can vary with the choice of function. In other words we try to match each of the peaks in $\boldsymbol{U}$ with the closest peaks in $\boldsymbol{R}$, starting with those closest to each other, continuing to match peaks with increasing distance until an upper limit d_{min} is reached. The result is a list $\boldsymbol{\Lambda} = \{q_l p_l\}$ where $l = 1,\ldots,L$ is the number of paired peaks, q and p, and $d_{q_1p_1} \leq d_{q_2p_2} \leq \cdots \leq d_{q_Lp_L}$.

The Accurate Mass Spectrum (AMS) Distance

When a correspondence between $\boldsymbol{U}$ and $\boldsymbol{R}$ has been established, we may evaluate the overall similarity between the spectra. Again, this can be done in several ways. One way is to ignore the absent peaks, and concentrate on the ones present and then apply the metrics described in the introduction as usual. Still, the differences in present and absent peaks are also descriptive, and have to be included in the evaluation.

The distance between $\boldsymbol{U}$ and $\boldsymbol{R}$ is a directional distance where

$$d_{U\mapsto R} = \frac{w_0}{L} \sum_{\{q_l,p_l\}\in\boldsymbol{\Lambda}} w\left(\phi^U_{q_l}, \phi^R_{p_l}\right) \cdot d_{q_lp_1} \tag{5}$$

which by definition may not fulfill the symmetry criteria of being a metric, since $d_{U\mapsto R}$ does not necessarily equal $d_{R\mapsto U}$, depending on the choice of the function d_{qlpl} and peak pair ratio w_0. Furthermore, $w(\phi^U_{q_l},\phi^R_{p_l}) \in [0;1]$ allow individual weights on each of the peak correspondences.

For $\boldsymbol{U}$ and $\boldsymbol{R}$ the global ratio of paired peaks w_0 can be given as

$$w_0 = 1 - \frac{L}{|\mathbf{U}|} \tag{6}$$

describing the fraction of peaks in $\boldsymbol{U}$ for which no mach was found in $\boldsymbol{R}$. The rationale is that a perfect match where all peaks in $\boldsymbol{U}$ are paired with peaks in $\boldsymbol{R}$ the match cannot better than w_0. Alternatively, if $w_0 = 1 - L/\max(|\mathbf{U}|,|\mathbf{R}|)$ is chosen, the similarity can never be larger than the ratio of paired peaks independent of direction.

The overall distance between the two spectra is calculated by combining the distances $d_{U\mapsto R}$ and $d_{R\mapsto U}$, through a maximum, minimum or by an average. [If eq 5 should be regarded as a true metric, it must fulfill the criteria of having $d_{U\mapsto R} = d_{R\mapsto U}$.]

As will be shown in the next section, this measure of distance can be used in a database search system, where $\boldsymbol{U}$ represents a query spectrum and $\boldsymbol{R}$ any given entry in the database.

More formally this is written

$$\phi^{S_k}_p = \left\{ m^{S_k}_p, i^{S_k}_p, w^{S_k}_p \right\} = \{m,i,w\}^{S_k}_p \tag{7}$$

where $p = 1,\ldots, |S_k|$ is the number of peaks in the k'th spectrum S_k. m is the peak mass, i the (log) intensity, and w is the width of the peak. Finally the k'th spectrum in the database can be written as

$$\mathbf{\Phi}^{S_k} = \left\{ \phi^{S_k}_p \right\} \tag{8}$$

The number of peaks, $|S_k|$, is allowed to vary across the library spectra, k.

Peak Similarity

To measure the distance d_{qp} between peak pairs in eq 5 we need a generalized distance metric, which takes peak width into account if spectra of various resolutions are to be compared. This can be done using the Jeffreys-Matusitas distance (JM-distance) providing a reliability criterion as a function of similarity between the peak pairs in two spectra based on a peak model [10].

The JM-distance is based on peak models, $g(m,\hat{\phi}_q)$ and $g(m,\hat{\phi}_q)$ for the peaks found in the spectra and can be described by

$$d_{qp} = J_{qp} = \left[\int_m \left(\sqrt{g(m,\hat{\phi}_q)} - \sqrt{g(m,\hat{\phi}_p)} \right)^2 dm \right]^{1/2} \tag{9}$$

To calculate the JM-distance we need a peak model, thus an estimate of the original ion population from which the centroid was calculated. No generalized mass peak models exist as the peak shape (peak width) depends on many factors including both instrument type and mass. A Gaussian peak model is sufficiently close for a generalized approximation

$$g(m,\hat{\phi}_p) = i_p e^{\frac{(m-m_p)^2}{2\sigma_p^2}} \tag{10}$$

where $\hat{\phi}_p$ is a centroid mass peak described by (see eq 3) the mass m_p, intensity i_p, and peak width (FWHM) w_p. The relation between the peak width (FWHM) w_p and σ_p is given by

$$\sigma_p = \frac{1}{2\sqrt{2\log 2}}\, w_p \simeq \frac{1}{2.3548}\, w_p \tag{11}$$

This peak model is not used to assess peak parameters or mass accuracy, rather it is used to estimate whether a pair of peaks belongs to the same or different population in the mass domain. If the absolute accuracy is known this can be used to limit search range in the peak pairing process, but it is not used in the similarity calculation. By this assumption, the peaks can be approximated by combining the Gaussian model eq 10 with the JM-distance (eq 11) to get (after normalization by square root)

$$d_{qp} = \sqrt{1 - e^{-\alpha_{qp}}}, \qquad d_{qp} \in [0;1] \tag{12}$$

where α_{qp} is called the Bhattacharyya distance [11, 12]. α_{qp} describes the standardized distance between peak p and q and based on **both** means (mass) and dispersions (resolution). The Bhattacharyya distance, α_{qp}, can be calculated as

$$\alpha^2_{qp} = \underbrace{\frac{1}{4}\frac{(m_q - m_p)^2}{\sigma_q^2 + \sigma_p^2}}_{\text{A}} + \underbrace{\frac{1}{2}\ln\frac{\sigma_q^2 + \sigma_p^2}{2\sqrt{\sigma_q\sigma_p}}}_{\text{B}} \tag{13}$$

The first term, A, describes the standardized distance between the centroid mass values, the latter term, B, express the difference in peak width. Eq 13 can be modified by normalizing the peaks to the same peak area (or height) and width thus $i_p = i_q = i_0$ and $\sigma_q = \sigma_p = \sigma_0$ both of which are constant. In this case eq 13 is reduced to describe the distance and will focus on qualitative features assessing only the mass difference between the two spectra.

Table 1. *Penicillium* species used in the study, with the corresponding notation

Label	Species	Number
A	*P. viridicatum*	7
B	*P. tricolor*	4
C	*P. aurantiocandidum*	17
D	*P. cyclopium*	8
E	*P. melanoconidium*	9
F	*P. polonicum*	9
G	*P. aurantiogriseum*	9
H	*P. neoechinulatum*	8
I	*P. freii*	9

All of the 80 isolates were taken from the IBT culture collection, held at biocentrum-DTU, Denmark, C and D have recently merged into one species by Frisvad and Samson [13].

Materials and Data Structure

As a part of an ongoing study of all known species in *Penicillium* and *Aspergillus*, about 500 isolates including 58 terverticillate *Penicillium* species sub-genus *Penicillium* were analyzed [13]. A sub-set of 80 isolates representing the nine closely related species in the series *Viridicata* was selected for development and illustration (see Table 1).

All isolates were taken from the IBT culture collection (BioCentrum–DTU, Kgs. Lyngby, Denmark) and inoculated in three points on czapek-yeast-autolysate agar (CYA) [14] and yeast-extract-sucrose agar (YES) [15] and incubated in the dark for 7 days at 25 °C. Cultures were identified by experts using all available phenotypic characters [13]. Extracts were prepared using the plug extraction procedure by Smedsgaard [16] using a two-step extraction with ethyl acetate containing 0.5% (vol/vol) formic acid in the first step and 2-propanol in the second step. The combined extracts were evaporated to dryness and re-dissolved in methanol. 1 μl methanol extract was infused from a FAMOS autosampler (LC-Packings Amsterdam, The Netherlands) using methanol as carrier at a rate of 15 μl/min. Just prior to the source, water containing 0.4% (vol/vol) formic acid was added through a T-piece at a rate of 5 μl/min giving a final combined flow of 20 μl/min with a composition of 75% methanol with 0.1% formic acid going into the ESI source.

Samples were analyzed on a Micromass Q-TOF system (Waters, UK) running MassLynx 3.5 with 3.6GHz Time to Digital Conversion (TDC). The instrument operated in positive electrospray and was tuned to a resolution better than 8500 FWHM and a 5th order mass scale calibration was made daily using a solution of polyethylene glycol 200, 400, and 600. Spectra were collected at 1 scan per s from *m/z* 150 to *m/z* 1000 in continuum mode. About 150 scans were collected from each sample each with about 118.000 data points.

The raw data were read directly from the MassLynx data files for processing by software written in-house.

After applying calibration, filtering, and internal mass correction (lock mass), the spectra were transformed into centroid data and stored in a flat database [9]. Each spectrum was stored with the following parameters: (accurate) mass, intensity, and resolution (FWHM), as illustrated in Figure 2. Prior to peak extraction, all the spectra were normalized to have a maximum abundance equal to one, in order to compensate for differences in concentration.

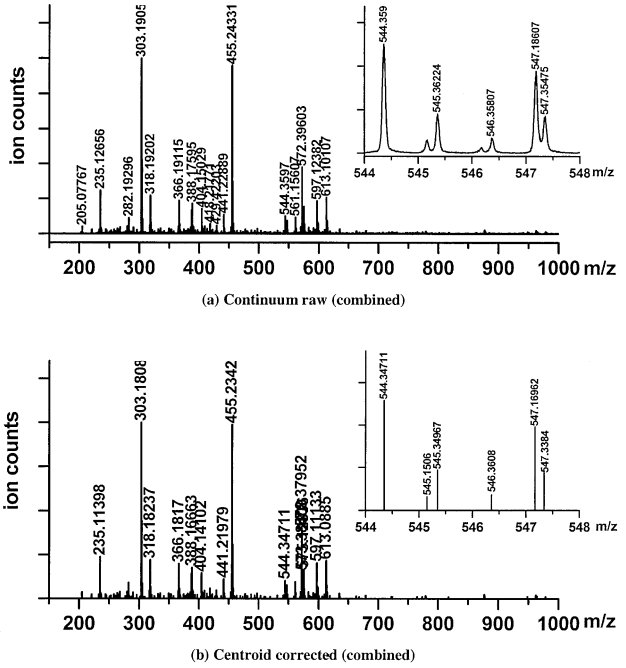

Figure 3. Example of a raw spectrum (upper) from the infusion of a crude extract and the corresponding centroid spectrum. From this raw spectrum the centroid mass spectra are calculated according to Hansen et al. [9] and internal mass correction is applied if an ion corresponding to a known metabolite is found obtaining an accuracy typical around 5 ppm.

Results and Discussion

The goal of a database search algorithm is to find similar mass spectra in a database. To do this in a qualified way the first goal is to collect a spectrum according to good laboratory practice and convert this into a suitable format [9]. Figure 3 shows an example of a raw continuum spectrum from the infusion of a crude extract along with the corresponding centroid spectrum. From these raw spectra the centroid mass spectra are calculated [9] and internal mass correction is applied if an ion corresponding to a known metabolite is found. In this process the peak width, thus the FWHM, is calculated for all ion peaks. There is no easy way to estimate the mass accuracy, although an estimate can be based on the residue of the external calibration and the mass correction calculated from the internal mass reference. Therefore, estimated accuracy is not used in the AMS library approach, however, it may be included in the future.

We then calculate the distance between all peaks in $\boldsymbol{U}$ with those in $\boldsymbol{R}$ to get $d_q. = \{d_{q1}, \ldots, d_{q|R|}\}$ which is the

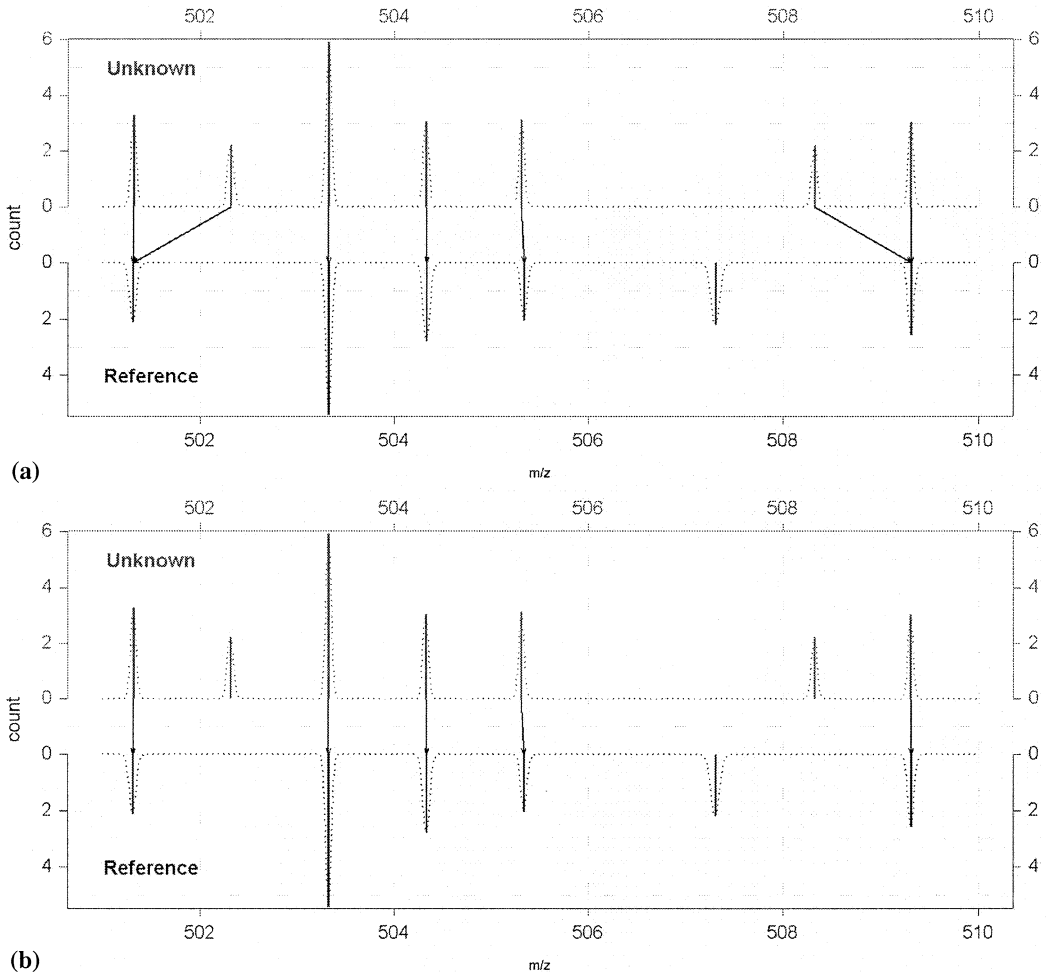

Figure 4. Correspondence between an unknown and a reference spectrum. (a) Correspondence by closest peak, (b) correspondence by unique closest peak. The spectra are plotted in the mass range of 501 Da to 510 Da. In both plots the dotted lines illustrate the estimated profile by the peak model of the original peak shape, and the vertical lines shown are centroid mass spectra used as input.

vector that describes the peak differences between the peak q in the unknown spectrum, and all peaks in the reference spectrum. Collecting all distances between peaks in $\boldsymbol{U}$ and $\boldsymbol{R}$ we get

$$\mathbf{D} = \begin{pmatrix} \mathbf{d}_{1\cdot} \\ \\ \mathbf{d}_{|U|\cdot} \end{pmatrix} = \begin{pmatrix} d_{11} & \cdots & d_{1|R|} \\ \vdots & \ddots & \vdots \\ d_{|U|1} & \cdots & d_{|U||R|} \end{pmatrix} \tag{14}$$

where $D \in \Re^{|U|\times|R|}$.

If we sort these distances to find the closest relatives between U and R in the mass domain we can establish a correspondence between peaks in the unknown spectrum to those in the reference spectrum as illustrated in Figure 4. Thereby we can build a unique list of peak correspondences where only one correspondence is allowed for each peak.

In general the distances are typical very low, in this case below ±0.01 Da. From the list of unique correspondences the Bhattacharyya distance between the reference and unknown spectra can be calculated by eq 13 to give the overall AMS distance.

To evaluate the performance of the new algorithm we compare the AMS performance to the algorithms described in [4]. From that study we have chosen the three best performing algorithms, because of their simplicity and because they are intuitively understand. These methods are all based on aligned data; we thus use a bin approach with varying bin width. A bin width of 1 Da corresponds to nominal mass spectra. In all cases the largest peak in each bin is selected. Also, the most efficient off-set is selected where possible, thus the 1 Da bins are selected as nominal mass −0.3 Da to nominal mass +0.7 Da.

To evaluate the results and compare different algorithms, we need a scheme for selection, based on similarities. By using the k-nearest neighbor selection criteria [17, 18], it is relatively easy to add and update information in a library. The performance results for each of the algorithms—expressed as percent of correct nearest neighbor identified—are listed in Table 2. Recall-reliability plots are sometimes used to document overall search system performance, were not used here as the principal measure of comparative performance because

Table 2. Results of library searching

	Size of bins				
	8.5 Da	4.25 Da	1 Da	0.5 Da	0.25 Da
Euclidian	0.8831	0.9093	0.9252	0.9688	0.9159
Absolute value	0.9217	0.9028	0.9282	0.9159	0.9282
Dot-product	0.9252	0.9375	0.9499	0.9746	0.9499
AMS			0.9746		

The performance is evaluated for different number of bins, over the mass range *m/z* 140 to *m/z* 1000.

they are not directly sensitive to the rank of retrieval and will depend on the absolute magnitude of distance values.

We see that the Accurate Mass Spectrum distance performance is comparable to the dot-product measure, but still have the major advantage that it is independent of how discrete the grid is.

In Figure 5 the distance matrices for CYA using the dot-product and AMS distance are compared. We see that the AMS distance forms more distinct clusters in each of the groups, except for *P. cyclopium* (D) that is separated into two groups. This is in full concordance with other findings whereas this study supports the division of *P. aurantiocandidum* (C) and *P. cyclopium* (D) into two species.

Conclusions

We have described a new matching algorithm specialized for accurate mass spectra. The method relies on robust detection of peaks in the mass spectrum. From each of the peaks descriptive statistics, i.e., mass, intensity and peak width is used to compare and establish a correspondence between peaks from pairs of spectra. After a correspondence has been established, an overall similarity between the spectra was found. The Jeffreys-Matusitas (JM) distance is used as the discriminative value between the peaks in the spectra to be compared, based on a simple peak model assumption.

Whereas the existing library search methods are bound to process binned variables, the AMS method presented here is independent of any binning alignment, and proven to perform equally well as existing methods [4].

If further information is available for each peak (e.g., by MS-MS or accuracy), the method described can be extended to include further these data modalities as well.

It has been the intention of the authors to describe the algorithm as a new instrumental tool. It is obvious that many parts of the algorithm can undergo speedup and optimization. Fast algorithms are crucial to fill the need for searchable databases containing accurate mass spectra. The resolution is used to access whether two ions belong to the same ion population or are actually two separate populations. This, of course, requires that the mass scale be correctly calibrated.

In most databases spectra are stored as centroid data thus reduced to mass-intensity pairs. However, we suggest that the resolution is estimated at the acquisition and included in the data base entries and if available also the mass accuracy. These parameters can be used to greatly improve the performance of library searches particularly when data originate from difference types of instruments.

Acknowledgments

The authors thank professor Jens Christian Frisvad (BioCentrum-DTU) for constructive discussions of as well as proofreading the manuscript. Ellen Kirstine Lyhne and Hanne Jacobsen are gratefully acknowledged for cutting the plugs, extraction, and analysis of the samples.

The project was supported by the Danish Technical Research Council under the project Programme for predictive biotechnology: Functional biodiversity in Penicillium and Aspergillus (grant no. 9901295).

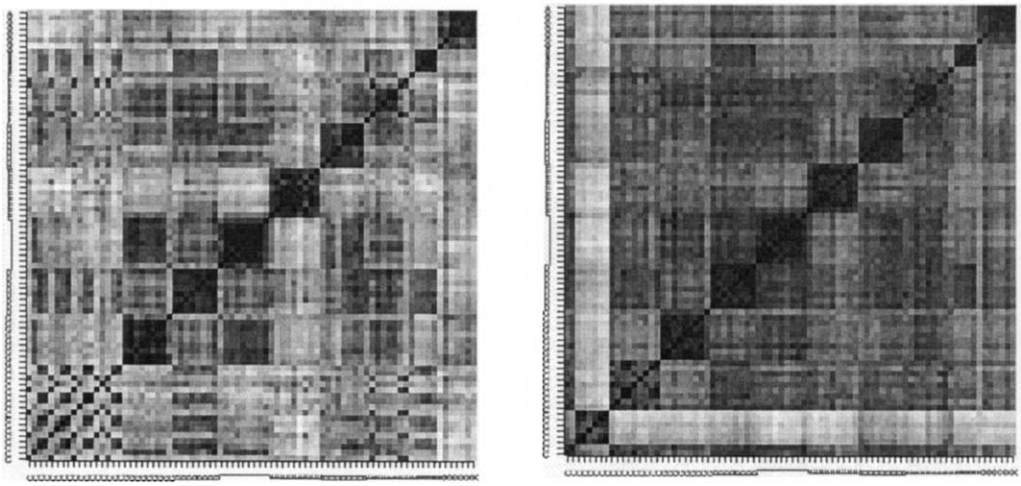

Figure 5. The distance matrix based on the dot-product similarity (left) and the AMS distance (right) from fungal extracts analysed by ES-MS. The fungi were grown on CYA agar.

Appendix A. Scripts and Benchmark Data

The statistical calculations have been made using "R", a language and environment for statistical computing and graphics. R is available as Free Software, and can be downloaded from www.r-project.org. The software used for extracting data from the MassLynx data files can be obtained together with a full documentation from www.metabolomics.dtu.dk or by contacting the corresponding author by email: meh@imm.dtu.dk. Furthermore we have made the data publicly available from this site, since the authors are of the opinion that having a benchmark data set is necessary in order to compare the performance of algorithms in the future.

References

1. Knock, B. A.; Smith, I. C.; Wright, D. E; Ridley, R. G. Compound Identification by Computer Matching of Low Resolution Mass Spectra. *Anal. Chem.* **1970,** *42(13),* 1516–1520.
2. Grotch, S. L. Computer Techniques for Identifying Low Resolution Mass Spectra. *Anal. Chem.* **1971,** *43(11),* 1362–1370.
3. Hertz, H. S.; Hites, R. A.; Biemann, K. Identification of Mass Spectra by Computer Searching a File of Known Spectra. *Anal. Chem.* **1971,** *43(6),* 681–691.
4. Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 859–866.
5. Stein, S. E. Chemical Substructure Identification by Mass Spectral Library Searching. *J. Am. Soc. Mass Spectrom.* **1995,** *6(8),* 644–655.
6. Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass Spectrom.* **2002,** *13(1),* 85–88.
7. Smedsgaard, J. Terverticillate Penicillia Studied by Direct Electrospray Mass Spectrometric Profiling of Crude Extracts: II. Database and Identification. *Biochem. Syst. Ecol.* **1997,** *25(1),* 65–71.
8. McLafferty, F. W. Probability Based Matching of Mass Spectra. *Org. Mass Spectrom.* **1974,** *9,* 690–702.
9. Hansen, M. E., Smedsgaard, J. On Automated Data Processing of High-Resolution Mass Spectra from Direct Infusion of Complex Fungal Extracts. *J. Am. Soc. Mass Spectrom.*, unpublished.
10. Matusita, K. Decision Rule, Based on the Distance, for the Classification Problem. *Annals Inst. Stat. Math.* **1956,** *8,* 67–77.
11. Fukanaga, K. Introduction to Pattern Recognition; *2nd ed;* Academic Press: Boston, 1990, p 103.
12. Gordon, A. D. *Classification*; *2nd ed;* Chapman and Hall: London, 1999, pp 15–29.
13. Frisvad, J. C., Samson, R. A. Polyphasic Approach of *Penicillium* subgenus *Penicillium*. A Guide to Identification of the Food and Airborne Terverticillate *Penicillia* and Their Mycotoxins. Studies in Mycology (Utrecht), in press.
14. Pitt, J. I. The Genus *Penicillium* and its Teleomorphic States *Eupencillium* and *Taleromyces*; Academic Press: London, 1979, pp 17–18.
15. Samson, R. A., Hoekstra, E. S., Frisvad, J. C., Filtenborg, O. *Introduction to Food and Airborne Fungi, 6th ed.*; Centraalbureau voor Schimmelcultures: Utrecht, 2000, p 379.
16. Smedsgaard, J. Micro-Scale Extraction Procedure for Standardized Screening of Fungal Metabolite Production in Cultures. *J. Chromatogr. A* **1997,** *760,* 264–270.
17. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning; Datamining, Inference and Prediction*; Springer, pp 378–382, 2002.
18. Roussopoulos N., Kelley S., Vincent F. "Nearest neighbor queries," in *Proceedings of the 1995 ACM-SIGMOD Intl. Conf. on Management of Data,* 1995, pp. 71–79.