Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification

Stephen E. Stein
NIST Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

Donald R. Scott
Atmospheric Research and Exposure Assessment Laboratory, U. S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

Five algorithms proposed in the literature for library search identification of unknown compounds from their low resolution mass spectra were optimized and tested by matching test spectra against reference spectra in the NIST–EPA–NIH Mass Spectral Database. The algorithms were probability-based matching (PBM), dot-product, Hertz et al. similarity index, Euclidean distance, and absolute value distance. The test set consisted of 12,592 alternate spectra of about 8000 compounds represented in the database. Most algorithms were optimized by varying their mass weighting and intensity scaling factors. Rank in the list of candidate compounds was used as the criterion for accuracy. The best performing algorithm (75% accuracy for rank 1) was the dot-product function that measures the cosine of the angle between spectra represented as vectors. Other methods in order of performance were the Euclidean distance (72%), absolute value distance (68%), PBM (65%), and Hertz et al. (64%). Intensity scaling and mass weighting were important in the optimized algorithms with the square root of the intensity scale nearly optimal and the square or cube the best mass weighting power. Several more complex schemes also were tested, but had little effect on the results. A modest improvement in the performance of the dot-product algorithm was made by adding a term that gave additional weight to relative peak intensities for spectra with many peaks in common. (J Am Soc Mass Spectrom 1994, 5, 859–866)

The present method of choice for the analysis of organic compounds in complex mixtures is low resolution, electron ionization mass spectrometry in combination with chromatographic separation techniques. Thousands of instruments with this capability are purchased each year, a large proportion of which are delivered with a computerized data system that contains a searchable reference library of mass spectra. Searching this library for compounds that have spectra that match a measured spectrum is the most common use of this library. This is usually done either to confirm a tentative identification or to suggest a list of candidate compounds that may have generated the measured spectrum. Compounds that have similar spectra are presented in a "hit list," which is ordered according to a numerical similarity index generated by the search algorithm. The details of the search algorithm can determine the effectiveness of the search [1],

and various algorithms have been proposed and implemented [2].

Clerc and co-workers [1, 3] have examined the general methods of searching spectral libraries. Two methods commonly applied in mass spectroscopy are identity searches and similarity searches; each has different requirements. In an identity search the spectrum of the unknown compound is presumed to be contained in the reference library and only experimental variability prevents a perfect match of unknown and reference spectra. In the more difficult similarity search the reference library does not contain a spectrum of the unknown compound and may not even contain similar spectra. Information that concerns the structure of the unknown must be inferred from the list of matches and their similarity indices. The present analysis is concerned only with identity searches.

An ideal identity search algorithm would assign probabilities to the candidate compounds in the hit list that reflect the true likelihood that each is the correct match. Although recent progress has been made in this area [4], inherent difficulties in determining the proba-

bility that an unknown compound is actually in the library require the use of empirical relations. Spectra of the same compound obtained on different instruments or under different analysis conditions can differ substantially. Even more serious is the inability of conventional mass spectrometry to distinguish clearly between members of certain classes of isomers and homologs. In these latter cases even the best matching algorithm along with the highest quality spectra cannot guarantee correct identification.

These factors will limit the performance of any search algorithm, and the relative performance of each algorithm is to a degree a measure of how successfully it deals with them. Determining retrieval accuracies is also straightforward because the compound identities in the test set are known. Such testing does, however, require the use of a large representative test set with searching against a single comprehensive reference library.

Two algorithms have gained widespread use due to their availability on commercial mass spectrometer data systems. One of these is the thoroughly documented probability-based matching system (PBM) of McLafferty and co-workers [5], which uses peak occurrence statistics in its spectral comparison logic. The other algorithm, first widely used in the INCOS data system* [6], compares unknown and reference spectra by computing the cosine of the angle (dot-product) between their vector representations. Performance of the latter algorithm has not been documented except in a limited environmental context [7] and details of its implementation appear only in a technical manual [6]. Rasmussen and Isenhour [8] performed the only previous comparative study of search algorithms for mass spectra with a relatively large database, but with a test set of only 40 spectra and without the use of the above-mentioned commercially implemented algorithms. The primary objective of the present study is to evaluate the retrieval performance for a wider range of algorithms with a much larger test set. Furthermore, when possible, algorithm performance is optimized by varying the scaling and mass weighting of the peak intensities. An optimal composite search algorithm, which used the optimized dot-product function together with an additional term based on ratios of peak intensities, also was developed from an analysis of performance results.

Background

A mass spectrum can be represented as a row vector composed of the ordered peak intensities in a spectrum. An individual spectrum also can be considered to represent a single point in a multidimensional hyperspace defined by the mass (m/z) variables considered. Each of the intensities in the row vector represents the value of the coordinate of the spectral point along the individual mass axis in this hyperspace. See Figures 1 and 2 for an example in three dimensions. If two spectra being compared are identical with respect to all the mass intensities, then their point representations in this hyperspace will coincide. If spectra are very similar with respect to the chosen mass variables, their point representations will be close in the space.

In spectral library searching the ideal situation would be to obtain a perfect match of the unknown spectrum with a single library spectrum. This would require one-to-one correspondence between each peak intensity in the unknown and reference spectra. Because of instrumental variability, this generally is not possible. Instead, the matching process can be visualized as the determination of the spectral points within
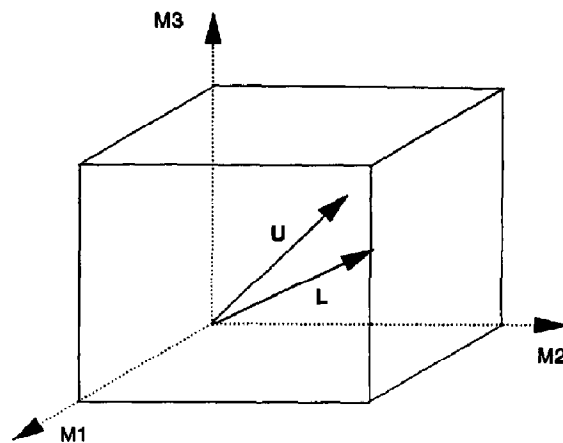

Figure 1. Vector representation of a hypothetical three-peak unknown (U) and library (L) mass spectrum in three-dimensional space (peaks have mass M1, M2, and M3).
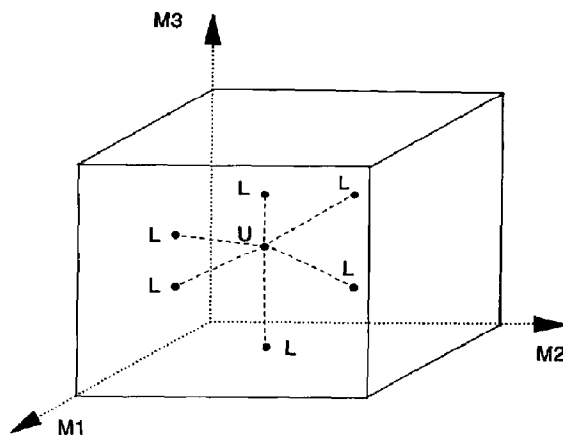

Figure 2. Point representation of library search results (L) for a hypothetical three-peak unknown (U) spectrum (masses M1, M2, and M3)

a specific volume element centered at the spectral point of the unknown in the hyperspace determined by the search variables. For the case where the unknown is present in the library, the hypervolume will be ideally contracted to include a unique point closest to the unknown spectral point. For a case where the unknown is not in the library, those spectral points within the defined hypervolume are listed up to a maximum distance and structural information concerning the unknown is inferred from this list. See Figure 2 for an example in a three-dimensional space. If a fixed number of spectral points are considered, this method is equivalent to the so-called K-nearest-neighbor method.

The various search algorithms that have been proposed consist of two major components that determine the differences among them: data preprocessing and spectral similarity calculation. Data preprocessing consists of peak selection, peak intensity scaling, and weighting of peak intensity by mass position. The similarity of two spectra, which may be regarded as the inverse of their "distance" apart, has been measured in various ways, including a Euclidean distance (sum of the squares of the differences between the chosen weighted peak intensities), an absolute value distance (sum of the absolute differences between the chosen weighted peak intensities), or some less easily interpreted "similarity index" such as the one used by Hertz et al. [9]. This latter index is an average weighted ratio of the known to unknown spectra taken mass for mass divided by 1 plus the fraction of intensities in both spectra that are not matched. The empirical weights for the ratios in the average are dependent on the larger of the intensities in the ratios. The dot-product algorithm [6], which is commonly employed, uses the cosine of the angle between the unknown and library spectral vectors. For unit-normalized spectral vectors, this angular measure is equal to 1 minus one-half the square of the distance between the two spectral points in the multidimensional space and therefore is also a distance measure. The PBM method, unlike the others, cannot be represented simply as an analytic function; it is based on peaks in an unknown and library spectrum falling within predefined abundance windows. Nonmatching peaks in the unknown spectrum are treated as impurities, unmatched peaks in the reference spectrum are "flagged," and both peak occurrence probabilities and empirical corrections are used to produce "reliability rankings" that serve to order the candidate list [5b].

Procedures

Library and Test Spectra

The 1992 release of the National Institute of Standards and Technology–Environmental Protection Agency–National Institutes of Health Mass Spectral Database, (hereafter called the NIST database) of 62,235 compounds served as the library of reference spectra for these studies. A collection of 12,592 alternate spectra of about 8000 compounds that were represented in the database was used as the test set. These spectra comprise the selected replicates file, which is provided along with the NIST database. Each compound in this set has a Chemical Abstracts Service registry number that matches one in the NIST database. All spectra in this collection have been selected from a larger archive on the basis of quality by an experienced evaluator. The selected spectra were contributed from a large number of laboratories and they represent a wide range of chemical structures and analytical conditions of practical interest. Another test set of 370 spectra, which had been used previously to optimize the performance of the PBM system [5b, d], were used principally to evaluate our implementation of PBM.

Search Software

A modified version of the software supplied with the personal computer version of the NIST database was used for library searching. Only "forward" searches were performed (all peaks in test and library spectra were used in matching). To improve the speed of retrieval, a single preliminary screening step was used with all algorithms. This selected those library spectra that had the largest number of major peaks in common with the test spectrum. The step accepted, on average, about 100 spectra per search and was 98% effective in retrieving the matching spectra. To test for any influence of screening on results, in separate studies all screened-out matching spectra were specifically added to those spectra that passed the screen. This had no effect on the relative search performance of the different algorithms.

Search Algorithms

In the implementation of most algorithms, spectral compression steps that may have been applied previously were not used. Compression had been done primarily to reduce computer storage requirements. This is no longer necessary. Also, no spectrum screening methods used in previous studies were employed here because the present screening algorithm, as mentioned above, was very efficient and produced no bias. Peak intensity weighting, where the contribution of a peak to the similarity computation was represented as [mass (m/z)]^n [abundance]^m, was adjusted for most algorithms to find optimal n and m values, as explained later. The only exception to the these adjustments was for the implementation of PBM, where recommended spectral compression and peak weightings were needed to use available correlation tables [5,10].

The algorithms examined are listed in Table 1. The Hertz et al. similarity index and the Euclidean distance and absolute value distance measures were previously studied by Rasmussen and Isenhour [8]. PBM and the dot-product algorithm have not been subjected to com-

Table 1. Search algorithms investigated[a]

Euclidean distance

$$\left(1 + \frac{\sum (W_L - W_U)^2}{\sum W_U^2}\right)^{-1}$$

Absolute Value Distance

$$\left(1 + \frac{\sum |W_L - W_U|}{\sum W_U}\right)^{-1}$$

Hertz et al. [9]

$$\frac{\text{average of weighted peak intensity ratios}}{[1 + \text{fraction of unmatched intensities}]}$$

Dot-product (cosine), F_d

$$\frac{\left(\sum W_L W_U\right)^2}{\sum W_L^2 \sum W_U^2}$$

Probability-based matching (PBM) [5b, 5d, 10c]

Uses probability that, by chance, peaks match within an abundance window (W value) by using uniqueness values for mass (U value) and abundance (A) along with a variety of rules and correlation tables.

Composite:

$$\frac{N_U F_D + N_{L\&U} F_R}{N_U + N_{L\&U}}$$

F_D = Dot-Product Term Above
F_R = Ratio of Peak Pairs (below)
W = [Peak Intensity]^n [Mass]^m = Weighted Intensity
N = Number of peaks

$$F_R = \frac{1}{N_{L\&U}} \sum_{i}^{L\&U} \left(\frac{W_{L,i}}{W_{L,i-1}} \frac{W_{U,i-1}}{W_{U,i}}\right)^n$$

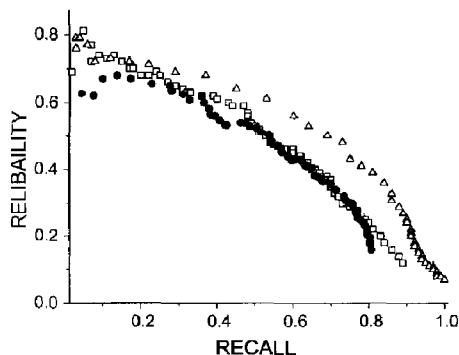where $n = 1$ or -1 when the term in parentheses is less than or greater than unity, respectively

[a] Subscripts: L = library. U = unknown, L & U = peak in both library and unknown spectrum.

Figure 3. Recall-reliability plot [13] for PBM and dot-product algorithms with 370 test spectra [11]. The dot-product (triangles) uses conventional peak weighting ([mass]^1 [abundance]$^{1/2}$); PBM, present implementation (squares); PBM results from ref 11 (circles) Recall is the fraction of all correct hits retrieved and reliability is the probability that a retrieval is correct.

parative examination. The composite method originates from the present work.

The PBM comparison logic was relatively complex [5, 10] and its implementation deserves special comment. All recently recommended features [5b, d] were implemented, including spectral compression, contamination correction, peak flagging, reliability ranking, and quadratic scaling. The correctness of our implementation was confirmed by the agreement of our results with those previously reported [11]. Results of tests using the same set of 370 test spectra against the same NIST database are shown in Figure 3. Also, by using this test set, the present PBM implementation gave an accuracy, counting all stereoisomers as correct hits, of 67% versus a reported 69% [12] against another comprehensive database. Furthermore, addition of several features recommended for improvement of PBM showed improvements in search results comparable in magnitude to those reported [10b, c].

Peak Intensity Scaling

Preprocessing of spectral intensities can have a beneficial effect on search performance and is commonly implemented in search systems. Peak intensities are frequently scaled to increase the relative significance of

smaller peaks. Because the higher mass peaks in a spectrum are the most diagnostic, weighting their contribution in the algorithm also improves performance. Both types of preprocessing were investigated in this study. Because of the complex interdependence of parameters in the PBM system and because they already have been optimized, no new scaling was applied to this system. However, "quadratic scaling" [5b] was used as recommended with PBM and also tested with other algorithms. This scaling procedure was designed to compensate for systematic mass–intensity variations due to mass discrimination by quadrupole mass filters or varying sample concentrations during scans. Two spectrum "normalization" schemes recommended in a commercial implementation of the dot-product algorithm [6] also were tested.

Performance Measurement

The primary measure of performance in this study was the rank of the correct compound in the hit list. This is easily understood and is independent of the numerical scoring schemes produced by the algorithms. It directly measures the key function of a search algorithm, to place the correct result as high as possible in the ranked candidate list. The correct library compound was taken as the one whose CAS registry number matched the test compound. Because each test compound had one spectrum in the primary NIST database, there was exactly one correct answer per test spectrum. For the purposes of testing, stereoisomers of test compounds in the library, even though they generally have indistinguishable mass spectra, were not counted as being correct.

Recall-reliability plots [13], which have been used to document overall search system performance, were not used here as the principal measure of comparative performance because they are not sensitive directly to

the rank of a retrieval and can depend on the absolute magnitude of similarity values.

Results and Discussion

Rasmussen and Isenhour [8], in their evaluation of mass spectral search algorithms, used a test set of 40 spectra with a library of about 17,000 spectra. They examined different peak intensity normalization schemes, various peak selection methods, and three distance or similarity measures. The two distance measures were the Euclidean and the absolute value. The Hertz et al. [9] similarity index also was studied. With this small test set this study did not discern statistically significant differences in search results for the algorithms, each of which made 29 to 30 correct rank 1 identifications. In another comparative test of methods of identifying mass spectra [14], where only environmentally important compounds were considered, the dot-product algorithm was found to give the best performance, exceeding even the accuracy of an experienced analyst. A test set of 85 compounds was used. Other methods evaluated in this latter study included a clipped Fourier transform spectral compression search system and a search algorithm based on a mass-weighted similarity index.

Algorithm Performance

A principal objective of the present study was to compare the search performance of the optimal forms of

the three algorithms studied by Rasmussen and Isenhour [8] as well as the more commonly used PBM and dot-product algorithms. Each algorithm was independently optimized for performance by using peak intensity scaling and peak weighting by mass, with the exception of the PBM system. The test spectra were a large set (12,592 spectra) of good quality replicate spectra. The compounds should be representative of practical mass spectral identification problems because the availability of a replicate spectrum generally means that the compound was of interest to more than one mass spectrometry laboratory.

The performance results, expressed as percent correct as a function of rank in the hit list, for the algorithms studied are given in Table 2. Some representative results obtained by varying mass and intensity scaling are included in this table along with the performance of the optimized algorithms. Table 2 also shows results obtained by using the scaling employed in a commercially available dot-product algorithm [6] (first-order mass power and square-root intensity). Comparing results for the optimized forms of the algorithms studied here, by using rank 1 results, shows that the dot-product algorithm performed best with 75% correct, followed by the Euclidean distance at 72% correct. The performances of the PBM and Hertz et al. algorithms were the worst at 64–65%. If ranks 1–3 are considered, the accuracies of all algorithms increased by approximately 18%. The results up to a rank of 10 are shown in Figure 4 for the most commonly used algorithms (dot-product and PBM) as well as for a composite algorithm that is discussed later.

Table 2. Performance of various algorithms

Algorithm	% Correct at Rank			Scaling /Comments	
	1	1 –2	1 –3	Mass Power	Intensity Power
Dot-product	72.9	85.9	90.8	1	0.5[a]
	73.2	86.3	91.0	1	0.5[b]
	72.8	85.9	90.8	1	0.5
Optimized	74.9	86.9	91.7	3	0.6
Euclidean distance	65.8	79.3	84.9	0	0.5
	69.9	82.9	88.2	1	0.5
Optimized	71.9	83.9	88.9	2	0.6
Absolute distance	61.4	74.9	81.2	0	1
	66.8	79.4	85.1	1	1
Optimized	67.9	80.3	85.5	2	0.9
PBM	57.1	71.5	78.5	"k value"[c]	
	64.0	77.7	84.3	Reliability[e]	
	64.7	78.4	84.8	Complete[e]	
Hertz et al.	59.9	73.9	81.1	0	See ref 9
Optimized	64.4	77.2	83.2	2	0.5
Composite	75.7	88.0	92.5	3, 0[f]	0.5, 1[f]

[a] Used local and global "normalization" [6].
[b] Used local "normalization" [6].
[c] Based on "reverse-search" overall spectral match factor [10].
[d] All recommended features except "quadratic scaling" [5b, d].
[e] All recommended features [5b, d].
[f] The first value is for the dot-product term and the second value is for the peak ratios term. The second and third power of the mass were equally effective for the first term.
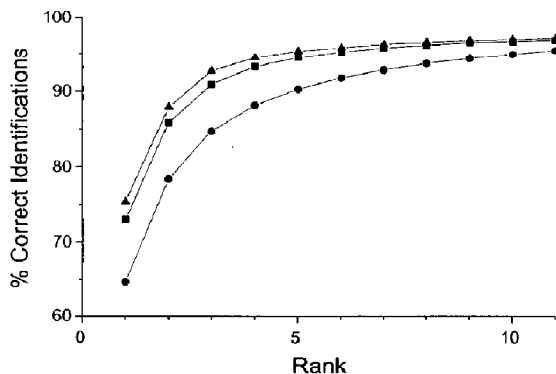
Figure 4. Search accuracy–rank position plot for composite, dot-product, and PBM algorithms. Composite (triangles); dot-product (squares) uses conventional peak weighting ([mass]^1 [abundance]$^{1/2}$); PBM, present implementation (circles). Each point gives the percentage of searches in which the correct retrieval was found among a given number of the best (top) retrievals.

Differences in performance are perhaps not as large as might have been expected considering the wide variety of comparison functions. This is due in large measure to the fact that a large fraction of the test spectra, perhaps 50%, are unique, so that virtually any simple distance measure will elevate the spectrum of a matching compound to the top of the hit list. In addition, some test compounds are mass-spectrometrically equivalent to other library compounds, so no algorithm could distinguish one from the other. These two factors, one that assists exact identification and the other that limits it, serve to significantly restrict the range of performance of any algorithm. This restriction was sufficiently strong to prevent the observation of significant differences in algorithm performance in the studies of Rasmussen and Isenhour [8]. Nevertheless, with the present large test set and optimized scaling, clear differences in the abilities of the different algorithms to place the correct match at the top of the hit list are evident. Examination of results shows that this is primarily due to their different abilities to find the one correct spectrum among a group of somewhat similar spectra.

No significant effects on the relative results would be expected if all mass-spectrally indistinguishable compounds were accepted as correct matches or if replicate spectra were present in the library. About 12% of the test compounds had stereoisomers or aromatic ring-positional isomers with virtually indistinguishable spectra in the library. If these were accepted as correct matches, the performance of all algorithms would have increased by about 5%.

Effects of Scaling and Mass Weighting

It is well known that scaling to compress the range of mass spectral peak intensities improves search perfor-

mance [15]. For most algorithms studied, scaling of intensities by the 0.5–0.6 power was optimal. (See Tables 2 and 3 for representative results.) Thus the commonly used scaling with the dot-product algorithm is nearly optimal. It also was found that logarithmic scaling of peak intensities, which has been used in interpretative analysis, drastically degrades retrieval performance because of exaggeration of the relative importance of low intensity peaks. Also, too low a power of scaling reduced the ability of spectral comparison functions to find the best match among library compounds that had similar, but distinguishable mass spectra. More complex intensity scaling procedures recommended to reduce effects of mass discrimination [6] were also tested with the dot-product algorithm. However, as shown in Table 2, they had no appreciable net effect on the search results. Any increase in the number of correct matches moved to the top of the hit list was balanced by the number of correct hits moved to lower positions. Although this form of scaling may offer some benefit for spectra known to be skewed, it does not seem to be needed for spectra obtained with properly calibrated spectrometers when using the present reference library. It should be noted that at the time this normalization procedure was first recommended, the intensity scale of quadrupole mass spectrometers often was not calibrated as it routinely is today.

It is common practice to place an increased weight on higher mass peaks. This deemphasizes the more variable and less characteristic lower mass range in a spectrum and emphasizes the more informative higher mass ions near the molecular ion. Weighting by various powers of the mass showed that squaring

Table 3. Effects of scaling on performance of the dot-product algorithm[a]

Scaling		Accuracy of Rank 1 Identification (%)
Mass Power	Intensity Power	
0	0.5	69.9
0.5	0.5	71.7
1	0.5	72.9[b]
2	0.5	73.6
3	0.5	73.7
4	0.5	72.6
1	0.3	62.8
1	0.4	69.4
1	0.5	72.9
1	0.6	73.5
1	0.7	73.2
1	0.8	71.9
1	0.9	70.9
3	0.6	74.9[c]

[a] These results are meant to be representative. Many other mass and intensity scaling values were tested on the dot-product algorithm. Other algorithms showed similar sensitivities.
[b] This is the scaling used most commonly [6].
[c] Optimized scaling; see Table 2.

or cubing was near optimal (Table 3). In an examination of simple exponential "tilting," the function exp(mass/50) was found to be near optimal, although it was less effective than the square or cube of the mass.

The more complex quadratic scaling, which is implemented in the PBM method, was tested with the other search algorithms, but had no net effect on their identification accuracy. Although useful in PBM for allowing the peaks of two spectra of the same compound to match within preset windows, such scaling appears to offer no advantage for the continuous comparison functions examined here.

PBM and Dot-Product Algorithm Performances

It is particularly interesting to compare the performance of the two commercially implemented algorithms—the complex statistically-based PBM algorithm and the dot-product algorithm. A clear difference in results of the two methods can be seen in Figure 3, which is a recall-reliability plot that used a set of 370 test spectra that had been used in PBM studies [11]. Results are shown for the original [11] and present implementation of PBM. A significant difference between PBM and dot-product results is their maximum recall, that is, the largest fraction of all correct hits retrieved. For the two PBM implementations this value is 81 and 89% versus 98% for the dot-product function. This means that 11–19% of the correct hits were rejected by PBM because not enough significant peaks fell within preset windows to pass its filters. This is the principal reason for the PBM recall-reliability curves falling below the dot-product curve at high recall.

Differences in performance also can be seen in Table 2, where the "conventional" dot-product algorithm produces 75% accuracy for rank 1 hits compared to 65% for PBM with the present large test set. This is also evident in Figure 4, where the percent accuracy is plotted against the rank on the retrieval list for the PBM, dot-product, and present modified dot-product (composite) algorithms. Both the dot-product and the composite algorithms outperform the PBM method even if the results up to rank 10 are used, although differences diminish with increasing rank. Approximately the same relative performance is exhibited using the PBM test set of 370 spectra (Figure 5). Examination of the search results suggests that these differences in performance are due to fundamental differences in matching criteria. The dot-product algorithm treats peak intensity differences in library and reference spectra in a continuous manner and produces its maximum value only when spectra are identical. PBM, on the other hand, uses abundance information only to determine whether or not peaks in the test and reference spectra fall into the same window. Although the use of narrower windows would improve the ability of PBM to find the best match among similar library
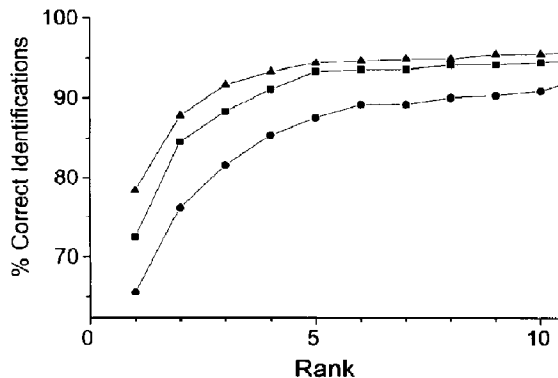

Figure 5. Search accuracy–rank position plot with 370 spectra test set [11]. Composite (triangles); dot-product with conventional scaling ([mass]^1 [abundance]$^{1/2}$) (squares); PBM, present implementation (circles). Each point gives the percentage of searches in which the correct retrieval was found among a given number of the best (top) retrievals.

spectra, this has been found to cause an undesirable reduction in its recall [10].

Composite Optimized Algorithm

As is apparent from Table 2, the dot-product algorithm is the best of the original five optimized search algorithms tested. Some effort was spent in an attempt to improve its performance. Manual inspection of search results in cases where the correct compound was not at rank 1 in the hit list, but where an experienced analyst might have selected it correctly, suggests that insufficient use of relative peak intensities for pairs of spectra with many peaks in common was a significant cause. Accordingly, an additional term was added to the optimized dot-product algorithm that measured only the relative intensities of neighboring peaks present in both library and test spectra (see Table 1). The contribution of this second term to the overall algorithm was weighted so that it increased in importance as the proportion of common peaks increased. This composite function, also listed in Table 1, outperformed the optimized dot-product algorithm alone as shown in Figures 4 and 5. Also, when the composite function was used, it was found that the search performance with first, second, and third power mass weighting became nearly equal.

Conclusions

Clear differences were found in the abilities of various optimized mass spectral comparison algorithms to correctly identify compounds. Of the previously proposed algorithms examined in the present study, the best performance (75% correct rank 1 identifications) was from the dot-product algorithm; next was the Euclidean distance method (72%), followed by the abso-

lute value distance method (68%), PBM (65%), and Hertz et al. (64%) methods. It is interesting that the two algorithms specially designed for mass spectrometry performed the worst. Various intensity normalization schemes proposed for correcting biased spectra had little effect on results. However, scaling peak intensities by taking their square root and weighting peak intensities by powers of the mass, usually a square or cube, did have a clearly beneficial effect. Based on attempts to correct deficiencies of the dot-product algorithm, a new composite algorithm was developed that modestly outperformed the others. It consisted of two terms. The first was a dot-product function with intensity scaling to the 0.6 power and weighting by the cube of the mass. The second term used nonscaled intensity ratios of neighboring peaks.

Rasmussen and Isenhour [8] reported no statistically significant differences in performance of three of the algorithms examined here. The present studies show an 8% difference in rank 1 identification for these algorithms. This apparent difference in results can be explained by the relatively small test set used in the earlier studies and differing peak processing and scaling procedures.

The search algorithm performances obtained in this study are realistic estimates with a diversity of test spectra. If one includes only the first candidate on the hit list, a 76% identification accuracy is expected. If the top three candidates on the hit list are considered, a 93% accuracy can be expected. The performance on complex real samples will certainly be less accurate. On the other hand, use of these algorithms with specialized libraries and higher quality unknown spectra is likely to produce better results.

The present work examined only the performance of algorithms for exact compound identification. Although features of "hit lists" produced by these algorithms may suggest that the unknown compound is not in the library [4, 5c], different algorithms are needed to optimally retrieve similar compounds. The testing and optimization of such "interpretive" algorithms is now underway in our laboratory.

Acknowledgment

References

1. Zurcher, M.; Clerc, J. T.; Farkas, M.; Pretsch, E. *Anal. Chim. Acta* **1988**, *206*, 161–172.
2. (a) Martinsen, D. P. *Appl. Spectrosc.* **1981**, *35*, 255–266; (b) Martinsen, D. P.; Song, B.-H. *Mass Spectrom. Rev.* **1985**, *4*, 461–490.
3. Clerc, J. T. In *Computer Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C., Isenhour, T. L., Eds.; Plenum Press: New York, 1980; pp 145–162.
4. Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 316–323.
5. (a) McLafferty, F. W.; Hertel, R. H.; Villwock, R. D. *Org. Mass Spectrom.* **1974**, *9*, 690–702; (b) Atwater, B. L., Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. *Anal. Chem* **1985**, *57*, 899–903; (c) McLafferty, F. W.; Stauffer, D. B. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 245–252; (d) Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 771–773.
6. Sokolow, S.; Karnofsky, J.; Gustafson, P. The Finnigan Library Search Program; *Finnigan Application Report* 2; Finnigan Corp.; San Jose, CA, March 1978.
7. Pellizzari, E. D.; Hartwell, T.; Crowder, J. A Comparative Evaluation of GC/MS Data Analysis Processing; *Project Report PB-85-125664*; U.S. Environmental Protection Agency: Research Triangle Park, NC, 1985.
8. Rasmussen, G. T.; Isenhour, T. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 179–186.
9. Hertz, H. S.; Hites, R. A.; Biemann, K. *Anal. Chem.* **1971**, *43*, 681–691.
10. (a) Pesya, G. M. Computerized Structure Retrieval and Interpretation of Mass Spectra: The Design and Evaluation of a Probability Based Matching System Using a Large Data Base; Doctoral Dissertation; Cornell University: Ithaca, NY, 1975; (b) Atwater, B. L. More Reliable Identifications of Unknown Mass Spectra Using the Probability Based Matching Algorithm; Doctoral Dissertation; Cornell University: Ithaca, NY 1980; (c) Stauffer, D. B. Improved Identification of Unknown Mass Spectra Using the Probability Based Matching Algorithm; Doctoral Dissertation; Cornell University: Ithaca, NY, 1984.
11. McLafferty, F. W.; Stauffer, D. B.; Loh, S. Y.; *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 438–440.
12. McLafferty, R. W.; Stauffer, D. B.; Twiss-Brooks, A. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 432–437.
13. McLafferty, F. W. *Anal. Chem.* **1977**, *49*, 1441–1443.
14. Lam, R. B.; Foulk, S. J.; Isenhour, T. L. *Anal. Chem.* **1981**, *53*, 1679–1684.
15. Crawford, L. R.; Morrison, J. D. *Anal. Chem.* **1968**, *10*, 1464–1469.