Estimating Probabilities of Correct Identification From Results of Mass Spectral Library Searches

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

This work presents a method for using mass spectral match factors reported by library search systems to obtain certain probabilistic indicators of correct identification. The overall probability that a retrieval is correct is formally separated into two independent terms. One of these is the probability that a retrieval is correct assuming that the correct match is contained in the library. This can be computed directly from test results. The other term represents the probability that the spectrum of the unknown compound is actually in the library. While the absolute value of this term cannot be computed, a relative value based solely on search results can be derived. This value may, if desired, be used to refine an initial estimate of the overall probability. Parameters used in this calculation are based on changes in test results caused by the logical removal of the test compounds from the library. These methods were parameterized from results of searching the NIST/EPA/NIH Mass Spectral Database with 12,592 good quality replicate spectra and a simple mass spectral comparison function. The methodology should be equally applicable to other libraries and search systems. *(J Am Soc Mass Spectrom 1994, 5, 316–323)*

Mass spectral library search systems are commonly used to help in the identification of unknown compounds from their electron impact spectra. In their most common application, these systems find and rank reference compounds whose spectra most closely match the spectrum of an unknown compound. Identification by this process relies on the simple concept that the more closely two spectra match, the more likely it is that they originated from the same compound. This process, variously called identity searching [1–3], straightforward searching [4], matching [5], or retrieval [5], uses a comparison function to assign values ("match factors") to reference spectra that provide a measure of their similarity to the spectrum of the unknown compound. These values are used to construct an ordered "hit list" that ranks the most similar library spectra according to their "distance from" the spectrum of the unknown compound. In practical applications, prior knowledge of the identity of the "unknown" compound ranges from none, in which case searching provides an initial list of candidates, to nearly certain, where searching is done to confirm a tentative identification.

The successful application of these methods requires that (1) a spectrum for the unknown compound is in the library and, (2) when the unknown compound is represented in the library, its spectrum is assigned a high match factor. The present work presents a means for computing probabilities associated with each of these factors. Estimates rely on results of test searches to provide correlations between match factors and probabilities of correct identification.

Available automated search systems make quantitative use only of *absolute* spectral match factors. Relative values reported in hit lists, on the other hand, while commonly considered by analysts for deciding whether a retrieval is correct, are not used in a quantitative way. In a method proposed by Biemann and co-workers [6], for instance, an asterisk is simply placed near the best hit if it is a sufficient distance from the second, reflecting an added degree of confidence in its correctness. The use of such differences for measuring the performance of search systems has also been discussed by Kwiatkowski and Riepe [7]. The present work describes a method that makes quantitative use of relative match factors, in addition to more generally used absolute match factors, for estimating factors underlying the overall probability that a retrieval is correct.

Procedure

Library and test spectra. The NIST/EPA/NIH Mass Spectral Database of 62,235 compounds [8] served as the reference library for these studies. A collection of 12,592 selected alternate spectra of approximately 8000 compounds, each with a CAS Registry Number, comprised the test set ("unknowns") for library searching.

1044-0305/94/$7.00

Received December 10, 1992
Revised November 10, 1993
Accepted November 10, 1993

Spectra in this file were selected from the Database source file by an evaluator based on quality [9]. Spectra were contributed by dozens of laboratories and span a wide range of compounds and analytical conditions. These compounds are broadly representative of those encountered in practice because the presence of a replicate spectrum in the source file usually indicates that the compound was of interest to more than one laboratory.

Search method. A modified version of the PC software distributed with the NIST/EPA/NIH Database was used for library searching. For rapid retrieval, a screening step selected for comparison only those library spectra with major peaks in common with the unknown. This step eliminated 5% of the correct retrievals. Searching with all 12,592 test spectra took 42 hours on a 33 Mhz, 386 PC.

Computed spectral match factors were derived from a weighted average of two comparison functions. The first is a measure of the "angle" between the two spectra [10], using scaling similar to that of the INCOS system [11]:

$$F_1 = \frac{\Sigma M(A_L A_U)^{1/2}}{[\Sigma M A_L \Sigma M A_U]^{1/2}}$$

For each peak, M is its mass-to-charge ratio value and A is its base-peak normalized abundance. Summations are overall peaks and L and U denote peaks in the library and unknown spectrum, respectively.

The second term is based on relative intensities of pairs of adjacent (nearest) peaks present in both spectra:

$$F_2 = \left(\frac{1}{N_{U\&L}}\right) \sum_{i=2}^{N_{U\&L}} \left(\frac{A_{L,i}}{A_{L,i-1}}\right)^n \left(\frac{A_{U,i}}{A_{U,i-1}}\right)^{-n}$$

$N_{U\&L}$ is the number of peaks common to the unknown and reference spectra, and $n = 1$ if the first abundance ratio is less than the second, or $n = -1$ if the reverse is true. The summation is over only those peaks present in both the unknown and library spectrum.

The match factor, MF, is obtained from F_1 and F_2 as follows.

$$MF = \frac{1000}{N_U + N_{U\&L}} (N_U F_1 + N_{U\&L} F_2)$$

Match factors range from 1000 for a perfect match to zero for spectra having no peaks in common. For each search, the twenty reference spectra with the highest match factors comprise the hit list, which is sorted by decreasing match factor. Rank denotes the position of a spectrum in this list (the best matching spectrum has rank 1) and $MF(i)$ is the spectral match factor for a hit of rank i.

Correct hits are generally defined here as those having the same CAS Registry Number as the test compound. Because the reference library has just one spectrum per CAS Registry Number, and all test compounds are also present in the NIST/EPA/NIH Database, there will be exactly one correct hit per search. In a separate set of searches, all stereoisomers of the test compound were accepted as correct hits (Class I matches [12]), thereby allowing, for some search compounds, more than one correct retrieval. Chemical structure processing software identified these stereoisomers.

An article is being prepared describing the performance of this algorithm, which is closely related to many of those used in commercial instruments. However, the general ideas presented here should be applicable to any comparison algorithm providing a quantitative measure of spectral similarity.

Results

Results needed to estimate the probability that a hit of rank r is correct, assuming the correct hit is somewhere in the library, $P_c(r)$, are discussed first. Results related to the likelihood that the unknown compound actually is in the library, $P_{present}$, are presented later.

Central to the present analysis is the strong correlation between the probability that a retrieval is correct and its *relative* match factor. This correlation is illustrated in Figure 1 for top-ranked hits, whose probabilities of being correct are plotted against their distance to the second best retrievals, $\Delta MF(1) = MF(1) - MF(2)$. Also shown are the numbers of searches having different $\Delta MF(1)$ values. Without using $\Delta MF(1)$, the probability of a top hit being correct could only be assigned a fixed probability of 0.74, but using these values provides more discriminating probabilities, with computed probabilities varying from 0.35 to over 0.9.

In contrast to the strong correlation between $P_c(1)$ and $\Delta MF(1)$, Figure 2 shows little relation between $P_c(1)$ and the absolute match factor, $MF(1)$. The decline in P_c at high MF values is due mainly to the presence

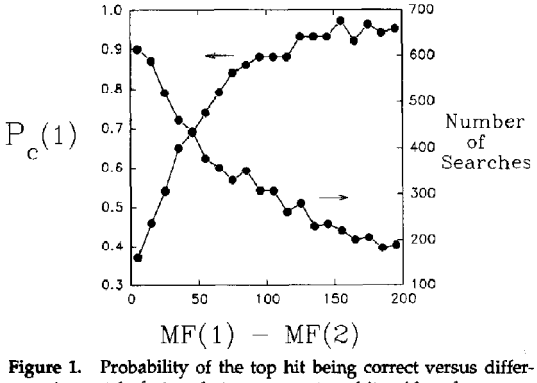

Figure 1. Probability of the top hit being correct versus difference in match factors between top two hits. Also shown are numbers of searches having various differences in match factors.

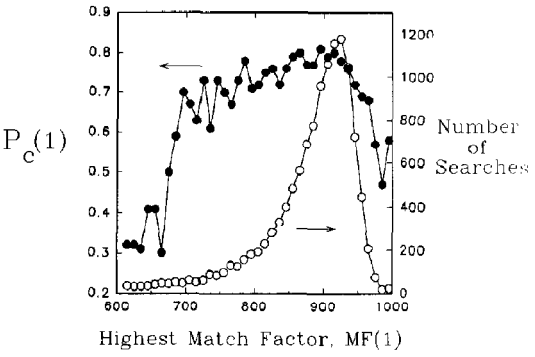

Figure 2. Probability of top hit being correct versus value of its match factor. Also shown are numbers of searches having various match factors.

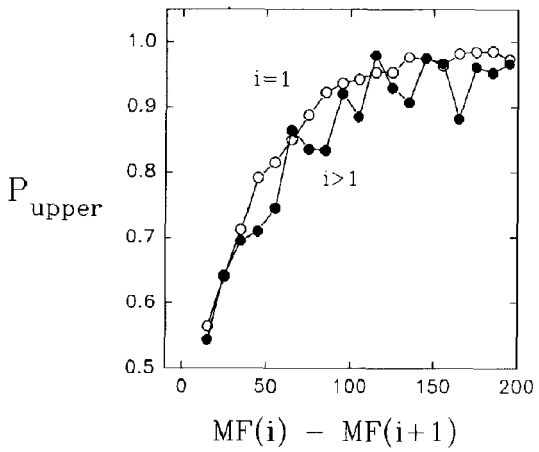

Figure 3. Likelihood that the upper (better matching) of a pair of adjacent retrievals is correct as a function of distance between their match factors ($P_{upper} = N_U/[N_U + N_L]$, where N_U and N_L are numbers of upper and lower correct retrievals, respectively). Open circles (rank $i = 1$) are for top two retrievals only; filled circles ($i > 1$) are for all others.

of an above-average proportion of simple aromatic ring-positional isomers having nearly identical spectra.

At first glance, the lack of correlation between $P_c(1)$ and $MF(1)$ might seem surprising because wrong first hits might be expected to have lower match factors than correct ones. However, a test spectrum will generally closely match the spectrum of the same compound in the library, so for an incorrect match to appear at the top position, it must match the unknown even more closely, albeit fortuitously. Top-ranked wrong hits therefore tend to have match factors at least as high as those typical of correct hits. An exception to this occurs when the correct retrieval is either lost in the screening step or is very different from the unknown spectrum and no similar spectrum is present in the library. This is the origin of the drop-off in $P_c(1)$ at lower MF values shown in Figure 2.

A more general expression of the close relation between the likelihood that a hit is correct and its relative match factors is presented in Figure 3. Shown here, as a function of distance between neighboring pairs of retrievals, is the relative likelihood that the upper of the pair is correct ($P_{upper} = N_u/[N_u + N_l]$, where N_u and N_l are numbers of correct upper and lower hits). Results for the top pair of retrievals are shown separately from all other pairs. The similarity of these two curves suggests that these probabilities are rank independent, so that the distance between two spectra is directly related to their relative likelihood of being correct whatever their position in the hit list.

The foregoing results concern the probability that a retrieval is correct with the implicit assumption that $P_{present} = 1$. We now derive values used later for dealing with $P_{present}$ itself. These values are derived from changes in search results caused by logically removing the spectrum of the matching compound from the library.

Two independent characteristics of hit lists were found to correlate with the presence of the unknown in the database. These are the absolute match factor of the top-ranked hit, $MF(1)$, and the largest difference in match factors of any two adjacent hits, ΔMF_{max}. Figure 4 shows the number of searches with various $MF(1)$ values when the unknown is in the database, N_p, and a separate curve shows results when the unknown is absent from the database, N_a. The latter case is simulated simply by ignoring correct hits in test searches. Likewise, values for N_p and N_a at different levels of ΔMF_{max} are presented in Figure 5. The relative independence of $MF(1)$ and ΔMF_{max} is reflected by correlation constants between these quantities of only 0.13 when the unknown compound is in the library and 0.14 when it is absent.

Effects on search results of accepting stereoisomers (Class I matches) as valid hits have also been examined. Approximately 10% of the test compounds had at least one stereoisomer in the library, leading to an increase of 15% in the total number of correct hits when using this criterion. Because of this relatively small proportion, the presence of stereoisomers had only a modest effect on the correlations discussed above, but accepting them as valid hits would increase the percent of correct top-ranked hits from 74% to 78%.

Discussion

Estimated Probabilities

This work presents a method for estimating certain factors underlying the probability that a compound retrieved in a mass spectral library search is correct. This is done by dividing the overall probability into two independent terms: (1) $P_{present}$, the probability that the unknown compound is in the database, and (2) $P_c(\text{rank})$, the probability that a retrieval of any rank (position in the hit list) is correct assuming $P_{present} = 1$.

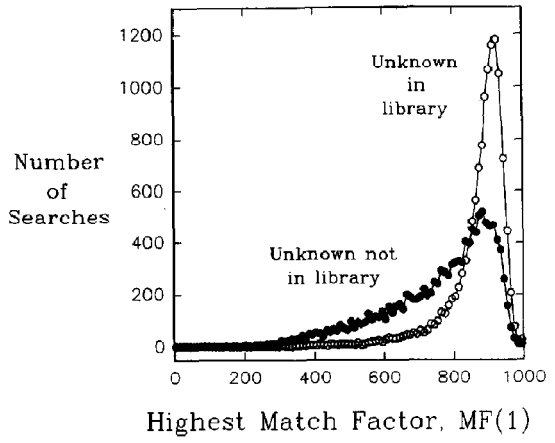

Figure 4. Number of searches at different top match factors with the correct match present in the library (open circles) and removed (filled circles) from the library.

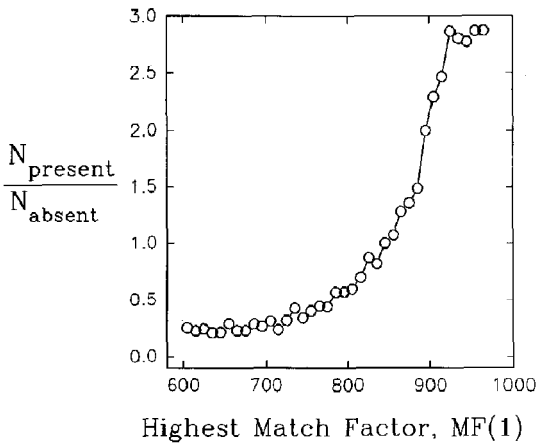

Figure 6. Relative number of hit lists having different top match factors when matching compound is present and absent from the library. This is the ratio of curves in Figure 4.

$P_{present}$

Because of the very large and uncertain number of compounds that a true unknown might be, it is not possible to derive a meaningful absolute value for $P_{present}$ solely from search results. On the other hand, certain "features" of hit lists can depend on whether the unknown compound has a spectrum in the library. Two such features have been identified: (1) the match factor of the best hit, $MF(1)$, and (2) the largest difference in match factors for adjacent retrievals in a hit list, ΔMF_{max}. We now present a means of using the dependence of these features on the presence of the unknown in the library, shown in Figures 4 and 5, to obtain a relative measure of the likelihood that a spectrum of the unknown compound is in the library, $R_{present,MS}$. It is then shown how to use this term to develop a formal expression for $P_{present}$.

$R_{present,MS}$ is obtained as follows. Because $N_p/\Sigma N_p$ is the probability of finding a given value of $MF(1)$ (or ΔMF_{max}) when the unknown is present in the library, and the corresponding probability when the unknown is not in the library is $N_a/\Sigma N_a$, then

$$R_{present,MS} = [N_p/\Sigma N_p]/[N_a/\Sigma N_a]$$

For the present studies, $\Sigma N_p = \Sigma N_a$, hence,

$$R_{present,MS} = N_p/N_a$$

which, for $MF(1)$ is the ordinate in Figure 6, and for ΔMF_{max} is the ordinate in Figure 7. The term $R_{present,MS}$ measures the degree to which hit list match factors

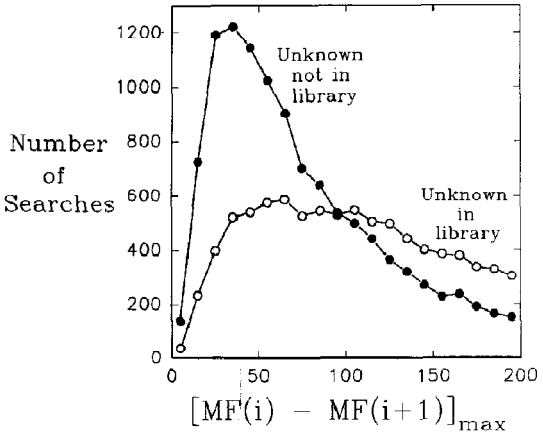

Figure 5. Number of searches at different maximum differences in match factors for adjacent retrievals with the correct match present in the library (open circles) and removed (closed circles) from the library.

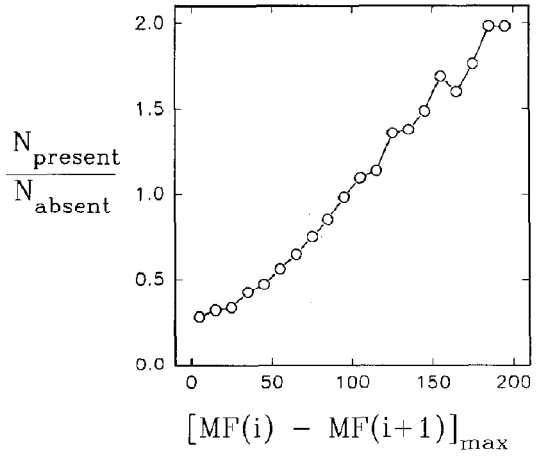

Figure 7. Relative number of searches having given maximum differences in match factors with matching compound present and absent from the library. This is the ratio of curves in Figure 5.

imply that a correct hit has been found. A value substantially greater than unity suggests that the unknown compound is in the hit list, hence in the library, while a value substantially lower than unity supports the opposite conclusion. While this term represents one of many factors that may be used by an analyst to decide if the unknown compound resides in the library, it is unique in that it may be derived solely from search results.

We now derive a relation between $R_{present,MS}$ and $P_{present}$. We first formally express the overall *relative* probability of an unknown being in the library as

$$R_{present} = R_{prior} R_{present,MS}$$

where R_{prior} is the relative probability, before consideration of library search results, that the unknown compound is in the library. Strictly speaking, R_{prior} must be supplied by the analyst (or an expert system) from information other than a mass spectrum. This may simply be a guess by the analyst based on previous experience for similar analyses. If no initial guess for R_{prior} is available, one may either arbitrarily assign it a value of unity (a 50% prior chance of being in the library) or simply report $R_{present,MS}$ to the user.

Next, a transformation from relative to absolute probability is needed:

$$R_{present} = P_{present}/(1 - P_{present})$$

The absolute probability is then derived as follows:

$$\begin{aligned} P_{present} &= R_{present}/(1 + R_{present}) \\ &= \left[1 + (1/R_{present})\right]^{-1} \\ &= \left[1 + 1/(R_{prior} R_{present,MS})\right]^{-1} \\ &= [1 + N_a/N_p R_{prior}]^{-1} \end{aligned}$$

Results for hit list features $MF(1)$ and ΔMF_{max} may be combined:

$$P_{present} = \left[1 + (N_a/N_p)_{MF(1)} (N_a/N_p)_{\Delta MF max} (1/R_{prior})\right]^{-1}.$$

Note that these ratios of N-values are inverses of values in Figures 6 and 7.

An important application of these probabilistic measures is their use as objective indicators of the need for *interpretive* search and analysis methods (as opposed to the present *retrieval* method). Interpretive methods are designed to find compounds with similar *chemical structures* rather than those with similar *spectra*. The STIRS [13] and SISCOM [14] systems are two well-known interpretive library search methods.

$P_c(rank)$

The following iterative procedure, illustrated in Table 1, uses the relative probability that the lower of two adjacent hits is correct, R_{lower} ($= [1 - P_{upper}]/P_{upper}$, using P_{upper} from Figure 3), to generate P_c for each member of a hit list.

$$P_{c,un}(1) = 1 \quad (1)$$

$$P_{c,un}(i) = P_{c,un}(i - 1) R_{lower}(\Delta MF(i)), \quad \text{for } i = 2 \text{ to } N. \quad (2)$$

$$P_c(i) = 0.945 P_{c,un}(i)/\Sigma P_{c,un}, \quad \text{for } i = 1 \text{ to } N. \quad (3)$$

$P_{c,un}$ is an intermediate, unnormalized probability and N is the number of reported hits (20 in most of the present searches). The value 0.945 in eq 3 is the fraction of all correct hits appearing in hit lists (of the 5.5% missing, 0.5% had ranks greater than N and, as mentioned above, 5% were lost in the screening step). A comparison of predicted and actual numbers of hits at each rank and at different levels of computed probability showed that this procedure worked well even at very low levels of predicted probability.

Note that while existing retrieval methods rely solely on absolute match factors for estimating probabilities of correct retrieval [15], the present scheme does not use them at all for this purpose. Instead, absolute MF values are used here solely to suggest whether the unknown compound is represented in the library.

Overall probability. The product of P_c for a given retrieval and $P_{present}$ for the search provides a formal measure of the likelihood that a retrieval is correct, $P_{overall}$. Whether it is preferable to actually use this quantity, or simply present P_c for each hit and $R_{present,MS}$ for the hit list, will depend on user requirements and knowledge. If there is some basis for estimating R_{prior} or if the user is comfortable with simply setting R_{prior} to unity, then $P_{overall}$ may be provided for each retrieval. This sort of analysis is in the spirit of Bayesian statistics [16]. Alternatively, one may simply report P_c and $R_{present,MS}$ to the user; these contain all of the derived statistical information, they are relatively straightforward to interpret and require only library search results as input.

Recall-Reliability Performance

McLafferty [17] has recommended that the performance of mass spectral library search systems be described by "recall/reliability" plots. Each point on these plots is derived from a set of retrieved spectra having spectral similarity values higher than a given value. Reliability is the fraction of members of this set that are correct, while recall is the fraction of all correct retrievals contained in the set. Figure 8 shows such a curve using P_c to define these sets and another curve that uses MF for this purpose. Clearly, the relative match factors used to obtain P_c possess a far greater ability to identify hits than do absolute match factors.

Also shown is a plot using the overall probability, $P_c P_{present}$ (with $R_{prior} = 1$), as the defining match factor. This curve describes search performance assuming a 50% a priori chance that the unknown compound is in the library.

Comparison to Probability Based Matching (PBM) Reliabilities

A different approach for using test search results to obtain probabilities of correct identification is implemented in the PBM search system [15]. This method uses individual match factors along with a variety of other terms to deduce "predicted reliabilities" [15]. Unlike the present method, these assignments can actually cause a reordering of compounds in the hit list. The values are intended to reflect probabilities that a retrieval and the unknown belong to the same class of compounds. The present approach interprets the hit list as a whole to provide relative probabilities that a given hit is the same compound as the unknown. In fact, these two approaches are quite complementary, and the predicted reliabilities from PBM could serve as input match factors in the present calculations. For this to be applied, however, correlations presented in Figures 3-5 would first have to be derived from PBM results.

Other Factors Affecting Estimated Probabilities

The present work employs a single reference library, a relatively simple spectral comparison function, and good quality test spectra. Further, only library compounds identical to the unknown are accepted as correct hits. We now examine the influence of these factors on estimated probabilities.

Table 1. Derivation of identification probabilities from match factors

Rank (r)	Match Factor (MF)	ΔMF[a]	P_{upper}[b]	R_{lower}[c]	$P_{c,un}$[d]	P_c[e]
1	850				1.0	0.48
		10	0.55	0.82		
2	840				0.82	0.40
		120	0.96	0.042		
3	720				0.034	0.016
		25	0.64	0.56		
4	695				0.019	0.0095
		10	0.55	0.82		
5	685				0.016	0.0077
		5	0.52	0.92		
6	680				0.015	0.0073
		5	0.52	0.92		
7	675				0.013	0.0063
		10	0.55	0.82		
8	665				0.011	0.0054
		0	0.50	1.00		
9	665				0.011	0.0054
		10	0.55	0.82		
10	655				0.009	0.0044

$\Sigma P_{c,un} = 1.94$
$(N_a/N_p)_{MF(1)}$[f] $= 1.0$
$(N_a/N_p)_{\Delta MF max}$[g] $= 0.67$
$P_{present}$[h] $= 0.60$
$P_{overall}(r)$[i] $= 0.60\ P_c(r)$

[a] $MF(r) - MF(r+1)$
[b] From smoothed $i = 1$ curve in Figure 3 using ΔMF in previous column.
[c] $[1 - P_{upper}]/P_{upper}$.
[d] Unnormalized probabilities from eq 2.
[e] Probability of being correct assuming matching compound is in database, using eq 3 and $\Delta P_{c,un} = 1.94$. For simplicity, hits 11-20 are not considered. A precise treatment would use $\Sigma_{c,un}$ values for these.
[f] Inverse of $(N_{present}/N_{absent})$ in Figure 6 at $MF(1) = 850$.
[g] Inverse of $(N_{present}/N_{absent})$ in Figure 7 at $\Delta MF_{max} = 120$.
[h] $[1 + (N_a/N_p)_{MF(1)}(N_a/N_p)_{\Delta FM max} R_{prior}]^{-1}$, assuming $R_{prior} = 1$, see text.
[i] $P_{present} P_c(r)$, this is the overall probability that a hit of rank r is correct.

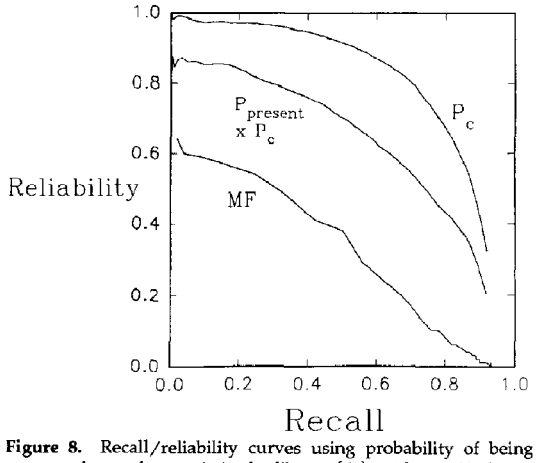

Figure 8. Recall/reliability curves using probability of being correct when unknown is in the library (P_c), with a 50% chance of the unknown being in the library ($P_c P_{present}$) and using only the absolute match factor (MF) assuming the unknown is in the library.

Reference library. The quality, size, and composition of the spectral library must affect retrieval statistics. High quality is especially important for compounds whose spectra are not highly unique, so that fine details can be used to elevate the correct match to the top of the hit list and separate it as far as possible from incorrect retrievals. A larger library will generally result in lower P_c values due to a greater chance of retrieving different compounds having spectra similar to the unknown. However, this may be offset by an increase in $P_{present}$ due to a more comprehensive coverage as long as a sizable fraction of the additional compounds are plausible candidates in some analyses. Smaller specialized libraries made up primarily of relevant compounds could lead to considerably higher values of P_c.

Comparison function. Numerical values from any comparison function that reflect the degree of similarity between the library and unknown spectrum can be processed by the methods discussed here. In fact, this method can be used with any similarity measure for complex "fingerprint patterns," including infrared spectra. However, results will generally be better when the best matching wrong hits are routinely reported, even if they have very low match factors.

The actual parameters used in the present calculations have, in effect, been calibrated for the present algorithm and library, hence they are not directly applicable to other search systems. The application of the present method, however, need not involve the very large number of test searches employed here. In fact, because the present method cannot produce highly accurate probabilities, we find that results of only several hundred searches can suffice for the development of practical systems. Assuming no dependence of $\Delta MF(i)$ on rank, as demonstrated for the present algorithm, R_{lower} may be derived from relative numbers of first and second rank hits at different $\Delta MF(1)$ values, using smoothing if necessary to derive P_c with eqs 1–3. $R_{present,MS}$ may be estimated from the same search results. Moreover, the general shapes of the curves derived from the present search results are not expected to depend drastically on a particular search system and may be used to guide any necessary extrapolations. In any case, because of unavoidable variations in spectral quality and compound class in practical applications, no model should be expected to be highly accurate.

Quality of unknown spectra. The present statistics were obtained by using essentially complete spectra that are free of major impurities, so that reported probabilities strictly apply only to searches with spectra of comparable quality. Spectra having major impurity peaks would have led to less discriminating measures of probability, although this might be partly offset by using a more appropriate comparison logic ("reverse searching" for example [18]) or spectral subtraction methods [19].

On the other hand, if searching principally involves thermally stable compounds and the instrument is properly tuned, actual search probabilities might be higher than those presented here. We find that the most significant source of legitimate variability between different classical electron impact spectra of a single compound arises from thermal decomposition before ionization. Also, many older replicate spectra used as test spectra came from, at least by today's standards, poorly tuned instruments.

Compound classes. The proper interpretation of library search results must consider the inability of mass spectroscopy to distinguish between certain classes of isomers and homologs. This has been considered in detail by McLafferty and co-workers [12], who distinguished four classes of compounds, each containing substances expected to have similar mass spectra. Class I, the narrowest class, includes all stereoisomers of a compound. Their spectra are generally indistinguishable. Class IV, the broadest class, includes various types of isomers and homologs known to have similar, though often distinguishable spectra. The use of any of these classes for identification implies that a single unknown spectrum may properly match more than one compound.

While the present results are based on there being at most one correct match in the library, extensions to cases where there are multiple correct matches are, in principle, straightforward. This may simply be done by either ignoring all but the best matching spectrum of a class or averaging their match factors. For Class I matching criteria we find that overall results differ little from those reported above because of the small proportion (15%) of additional correct hits, so the

present parameters can be used with little loss in accuracy in this case. Searches of libraries containing more than one spectrum of a compound may be treated in the same way (i.e., all but the best match are ignored or all match factors are averaged). Effects will be more significant for broader classes (Class IV, for example) where a large number of retrievals, perhaps the entire hit list, may be correct. However, the breadth of such classes limits their utility for identifying unknown compounds and presents a problem in implementing or even testing such methods on automated systems.

Other Correlations

Finally, we note that it may be possible to improve the discriminating power of the present method by using additional factors that may be independently correlated with P_c and $R_{present,MS}$. Some possible factors are the presence of a molecular ion peak in the library spectrum, the molecular weight of the reference compound, and the agreement between formulas for adjacent hits. Indeed, some of these are used to compute "reliabilities" in the PBM search system [5].

Summary

Procedures are described for using match factors reported by mass spectral library search systems to estimate certain probabilities that underlie the probability that a library retrieval is correct. These probabilistic terms can assist analysts in deciding which, if any, of the compounds retrieved by a library search match the compound that generated the submitted spectrum. These terms are computed by using parameters derived from results of a large number of test searches, and make use of both relative and absolute spectral similarity values (match factors). While reported parameters pertain only to the mass spectral comparison function examined here, the methodology can be applied to any search algorithm that provides a quantitative measure of similarity for submitted and library spectra. In fact, the general procedures described can be applied to similarity-based library retrieval systems for any type of spectra.

To assist in the interpretation of results, the overall probability for correct identification is formally separated into two independent terms: (1) the probability that the unknown compound is in the library and (2) the probability that a hit is correct assuming that the matching compound is in the library. The latter term can be derived directly from test results by a simple iterative procedure. The former term, however, cannot be derived from search results. Instead, a component of it may be derived that can serve several purposes. First, it can reinforce the judgment of an analyst as to whether a correct identification has been made. Second, it provides an objective measure of whether other means are needed to identify the unknown, including "interpretive" library searching. Finally, if an initial estimate of the probability that the unknown compound is in the library can be made, it can be used with search results and the present correlations to generate the probability that any reported retrieval is correct.

Acknowledgments

This work was supported in part by the U.S. Environmental Protection Agency under Interagency Agreement DW 13934923 to the National Institute of Standards and Technology. Technical discussions with D. R. Scott of the EPA are gratefully acknowledged.

References

1. Clerc, J. T. In *Computer Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C.; Isenhour, T. L., Eds. Plenum: New York, 1987; pp 145–162.
2. Scott, D. R. *Chemom. Intel. Lab. Sys.* **1988**, *4*, 47–63.
3. Domokos, L.; Henneberg, D.; Weimann, B. *Anal. Chim. Acta* **1984**, *165*, 61–74.
4. Cleij, P.; van'T Klooster, H. A.; van Houwelingen, J. C. *Anal. Chim. Acta* **1983**, *150*, 23–36.
5. McLafferty, F. W.; Stauffer, D. B. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 245–252.
6. Hertz, H. S.; Hites, R. A.; Biemann, K. *Anal. Chem.* **1971**, *43*, 681–691.
7. Kwiatkowski, J.; Riepe, W. *Fresnius Z. Anal. Chem.* **1980**, *301*, 300–303.
8. The NIST/EPA/NIH Mass Spectral Database. NIST SRD Database No. 1A, Gaithersburg, MD, 1992.
9. Stein, S. E.; Ausloos, P.; Lias, S. G. *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 441–443.
10. Rosenthal, D.; Hargrove, W. F. *Proceedings of the Annual Conference on Mass Spectrometry and Allied Topics*, 1979, 266–267.
11. Sokolow, S.; Karnofsky, J.; Gustafson, P. *The Finnigan Library Search Program.* Application Report Number 2, Finnigan Instruments, July 1978.
12. Pesya, G. M.; Venkataraghavan, R.; Dayringer, H. G.; McLafferty, F. W. *Anal. Chem.* **1976**, *48*, 1362–1368.
13. Haraki, K. S.; Venkataraghavan, R.; McLafferty, F. W. *Anal. Chem.* **1981**, *53*, 386–392.
14. Damen, H.; Henneberg, D.; Weimann, B. *Anal. Chim. Acta* **1978**, *103*, 289–302.
15. Atwater, B. L.; Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 899–903.
16. Casella, G. *Chemom. Intel. Lab. Sys.* **1992**, *16*, 107.
17. McLafferty, F. W. *Anal. Chem.* **1977**, *49*, 1441–1443.
18. Abramson F. P. *Anal. Chem.* **1975**, *47*, 44–48.
19. Atwater, B. L.; Venkataraghavan, R.; McLafferty, F. W. *Anal. Chem.* **1979**, *51*, 1945–1949.