Chemical Substructure Identification by Mass Spectral Library Searching

Stephen E. Stein
NIST Mass Spectrometry Data Center, Gaithersburg, Maryland, USA

A library-search procedure that identifies structural features of an unknown compound from its electron-ionization mass spectrum is described. Like other methods, this procedure first retrieves library compounds whose spectra are most similar to the spectrum of an unknown compound. It then deduces structural features of the unknown compound from the chemical structures of the retrievals. Unlike other methods, the significance of each retrieved spectrum is weighted according to its similarity to the spectrum of the unknown compound. Also, a "peaks-in-common" screening step serves to reduce search times and an optimized dot product function provides the match factor. If the molecular weight of the unknown compound is provided, the identification of certain substructures can be improved by including "neutral loss" peaks. Correlations between the presence of a substructure in a test compound and its presence among library retrievals were derived from the results of searching the NIST/EPA/NIH reference library with a 7891 compound test set. These correlations allow the estimation of probabilities of substructure occurrence and absence in an unknown compound from the results of a library search. This method may be viewed as an optimization of the "K-nearest neighbor" method of Isenhour and co-workers, with improvements that arise from spectrum screening, peak scaling, an optimal distance measure, a relative-distance weighting scheme, and a larger reference library. *(J Am Soc Mass Spectrom 1995, 6, 644–655)*

Mass spectral library searching is commonly employed to assist in the task of identification of unknown compounds. Widely available "identification" methods provide a "hit list" of compounds in a reference library whose spectra most closely match a submitted unknown spectrum [1–3]. These methods are designed to identify compounds represented in the library that might have generated the submitted spectrum, allowing for instrument-dependent variations of mass spectra. However, when the unknown compound is not in the library, these methods are less useful. Although nonidentical, but structurally similar compounds can appear in the hit list, most identification systems are not optimized to find them, and deriving reliable substructural information is not straightforward. In many cases, no similar spectra are retrieved, which indicates to the user only that the unknown compound is not in the library.

"Interpretive" library search systems [2–5], on the other hand, are designed to produce structural information for compounds not represented in the reference library. These methods typically employ a predefined set of spectral "features" designed to correlate with the presence of chemical substructures. Searching identifies the library spectra that have features most similar to those of the unknown spectrum. The frequency of occurrence of a substructure in the hit list is then used to estimate the probability that it is present in the unknown compound.

The identification of a substructure from a given mass spectrum can be difficult or even impossible, because its effect will depend on relative rates of competitive processes that depend, in turn, on other structural features of the molecule. Even for substructures that commonly produce characteristic patterns, actual "signatures" can be highly variable. Library searching deals with this variability by comparison of the submitted unknown spectrum to spectra of reference compounds that contain each substructure in a variety of chemical "environments." Substructures embedded in similar environments have an increased chance of having spectral features in common. Another virtue of the library search procedures in common use is that because they are derivatives of "K-nearest neighbor" methods, which have been shown to be especially effective for chemical classification [6–8].

Two well-developed interpretive search systems are SISCOM* [9, 10] and STIRS [11–13]. The reference compounds that produce library spectra most similar to an unknown spectrum, as measured by predefined

Address reprint requests to Dr. Stephen E. Stein, NIST Mass Spectrometry Data Center, National Institute of Standards and Technology, Receiving Room, Bldg. 301, Rt. 270 and Quince Orchard Road, Gaithersburg, MD 20899.

The mention of commercial products in this work does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that it is the best available product for the purpose.

1044-0305/95/$9.50
SSDI 1044-0305(95)00291-K

Received February 14, 1995
Revised March 31, 1995
Accepted April 4, 1995

features, comprise the hit list. Structures of these compounds are then analyzed to deduce whether a structural feature is present or, in a preliminary report of SISCOM extensions [14], absent in the unknown compound.

At the present time, a large proportion of modern mass spectrometer data systems allow library searching for compound identification. In this work, we describe and analyze an interpretive search system well suited for use alongside these systems. Not only does this system use algorithms and libraries similar to those commonly used for compound identification, but it is relatively easy to implement and relies on match factors already familiar to many users. It also can generate readily understood substructure present and absent probabilities for a wide range of structural groups. These probabilities may be directly used by the analyst for structure elucidation or may serve as input to automated structure generators [15, 16].

Background

It is well known that certain chemical substructures often reveal themselves as characteristic "signatures" in electron-ionization mass spectra. However, owing to the complex and often subtle relationships between the structure of a compound and its mass spectrum, the actual effect of a substructure on a spectrum can be hard to predict. As a consequence, the prediction of a complete spectrum from a structure is not generally possible, and large libraries of reference spectra are in common use. Gasteiger et al. [17] have, however, reported some recent progress in the prediction of mass spectra.

The reverse process, that of deducing the structure of a compound from its spectrum, is even more difficult. In fact, because many structural features have no clear signature, spectra for most compounds do not contain sufficient information for this purpose. However, mass spectra do commonly contain enough information to identify certain structural features with high reliability. A quantitative measure of the ability of a system to perceive a substructure is often expressed in terms of "recall." Recall is the percent of all compounds that contain a given substructure for which that substructure can be identified above a prespecified level of confidence. Recall depends on both the uniqueness of the "signature" of the substructure and the capability of the analysis system to perceive it.

Library searching has been shown to be an effective method to derive structural information from mass spectra. It avoids the need for a complex set of rules that relate spectra to substructures by, instead, comparison of the input spectrum to spectra in a comprehensive library that contain these substructures in a diversity of chemical surroundings. If characteristic features of a substructure in the unknown compound are revealed both in its spectrum and in some library spectra, search systems are responsible for finding these library spectra and placing them prominently in the hit list.

SISCOM, developed by Henneberg and co-workers [9, 10], first eliminates C-13 isotopic peaks and other less significant peaks within peak series, and then builds an initial hit list of library spectra that have the most features in common with the unknown spectrum. Peak intensities are then used to order the hit list. Although this system was initially designed to find similar compounds, it was later modified to identify identical compounds also [10]. Details of SISCOM algorithms and performance figures have not been published.

STIRS has been described in a series of publications and dissertations from McLafferty's group at Cornell [11–13, 18]. A variety of match factors are computed for each library and unknown spectrum pair and a separate hit list is maintained for each factor. Three general classes of spectral data are used for defining match factors—peak series, characteristic ions, and neutral losses. The use of the neutral losses requires advance knowledge of the molecular weight of the unknown compound. Following the creation of hit lists, a single match factor is derived from a linear combination of individual match factors and then used to build a single overall hit list. The number of compounds that contain each substructure among the top 15 retrievals are then counted. If, for any substructure, it is sufficiently unlikely that its number of occurrences in the hit list could have arisen by chance, then it is reported as being present in the unknown. No screening method has been reported for STIRS, and match factors are used only to find the top 15 retrievals. Also, substructure-absent predictions are not made. Considerable effort has been devoted to finding a comprehensive set of substructures for use by STIRS.

The K-nearest neighbor (KNN) method, as implemented by Isenhour and co-workers [7, 8], uses Euclidean distance to identify the "K" library spectra closest to the unknown spectrum. When a prespecified number of these library compounds contain (or do not contain) a substructure, the substructure is indicated as being present (or not present) in the unknown compound.

Curry and Rumelhart [19] recently described a promising new approach that uses library spectra to train an artificial neural network (MSNet). Results were comparable in accuracy to STIRS.

In a comparative study, Varmuza and co-workers [20] tested the ability of a number of different substructure identification methods to classify compounds from their mass spectra. It was concluded that for most substructures, mass spectrometry was not capable of providing reliable yes/no classifications, and that library searching was necessarily the best method for this purpose. These studies make it clear that if highly reliable identifications are required, classification systems must allow a "no-decision" result.

Method

Reference and Test Spectra

Spectra for the 61,500 compounds represented by chemical structures in the 1992 version of the NIST/EPA/NIH Database [21] served as the reference library. The test set was composed of a single spectrum for each of the 7891 compounds with structures in the NIST Selected Replicates Library. These test compounds are generally good quality, alternative spectra that broadly represent compounds of general interest. Retrievals of compounds identical to test compounds, as identified by matching Chemical Abstracts Service registry numbers, were omitted from hit lists.

Computer System

All library searching and analysis was done on an MS-DOS personal computer equipped with a 66-MHz Intel 486DX2 processor with 16-MB RAM. A modified PC version of the NIST/EPA/NIH Database [21], written in C, was used to search. The average search time was approximately 10 s.

Search System

The analysis of an input mass spectrum is done in three stages: screening, match factor calculation, and substructure probability estimation.

Screening. In principle, a library search should involve the direct comparison of the submitted unknown spectrum to each spectrum in the library. However, efficiency can be greatly increased by first rapid identification of a subset of library spectra with some similarity to the unknown spectrum and then comparison of only the spectra in this subset to the submitted spectrum. For this purpose, the present system uses a modified "ranked peaks-in-common" procedure to identify library spectra that have the most abundant peaks in common with the unknown. The INCOS identification search system [22] employs a similar procedure, called the "presearch." Library spectra that have the largest numbers of major peaks in common with the unknown spectrum, consistent with a specified minimum number of spectra to be retrieved, are rapidly identified in presorted files. The present system merges four subsets of screened spectra, and each has a different set of the following three requirements: (a) number of largest peaks in the unknown spectrum, (b) number of largest peaks in the library spectrum, and (c) minimum number of library retrievals. Specifically, the sets of these requirements (a, b, c) for each of these four subsets of screened spectra are: (1) 8, 15, 40; (2) 14, 14, 50; (3) 6, 6, 20; (4) 8, 8, 50. These values were derived for compound identification searching; no attempt was made to optimize them for the present substructure identification searching. It was found, however, that fine details had no effect on performance. These requirements led to an average of about 450 spectra per search. Overall identification accuracy was not measurably improved by increasing this number by a factor of 2.

When neutral losses were used to search, screening employed up to five of the largest loss peaks within m/z 72 of the molecular ion and abundances greater than 3% of the base peak. These values were found to be near optimal in trial runs, although further optimization might be possible. For a library loss peak to match a corresponding peak in the unknown spectrum, their abundances were required to be within a factor of 5 of each other.

Match factors. The normalized dot product of unknown and library spectra provides the basic measure of spectral similarity:

$$\mathrm{MF} = 1000\frac{(\Sigma A_U^{1/2}A_L^{1/2})^2}{\Sigma A_U \Sigma A_L}$$

Sums are over all peaks (mass-to-charge ratio values), and A_U and A_L are abundances for unknown (test) spectra and library spectra, respectively. The 25 best matching library spectra, ordered by decreasing match factor, comprise the hit list. Tests showed that the square-root abundance scaling used in this expression was near optimal, similar to that found in earlier compound identification studies [3]. However, unlike the earlier studies, no improvement in performance was gained by weighting peaks by their mass-to-charge ratio values. Substructure identification, unlike compound identification, makes effective use of both low and high mass peaks [23].

Probability estimator. Match factors of retrievals relative to the highest match factor in the hit list are the principal quantities used to estimate probabilities of substructure occurrence. The weight of a retrieval of rank r (rth member of a hit list) is

$$2^{(\mathrm{MF}(r)-\mathrm{MF}(1))/75}/N$$

where MF(r) is the match factor of the retrieval of rank r and N is the hit list normalization factor,

$$\sum_{r=1}^{25} 2^{(\mathrm{MF}(r)-\mathrm{MF}(1))/75}$$

This functional form and the scaling factor of 75 were found to be optimal by trial and error. The overall

weight of substructure s is W_s, the fractional weight of retrievals that contain the substructure:

$$W_s = \frac{\Sigma_{\text{has } s}\, 2^{(\text{MF}(r)-\text{MF}(1))/75}}{N}$$

Correlations between W_s and substructure occurrence in test compounds were derived from search results.

Separate tests showed that a similar scheme in which retrievals were weighted by their absolute match factors was not as effective as the present relative value weighting scheme. Also, as demonstrated later, assigning an equal weight to each of a fixed number of top hits was less effective than the present scheme.

Substructure Definition

The term "substructure" is broadly defined here to include any feature that can be derived from a molecular structure. Most substructures used here were taken from two lists developed for use with STIRS. One of these lists consisted of substructures that generally correspond to simple functional groups [12]. However, these substructures were derived from a linear structural representation, so special care had to be taken to reproduce the original substructure definitions. The subset of these substructures used in earlier comparative studies [8] also is used here. The most common substructures in a later, more comprehensive list [13] are also examined in the present work.

Molecular Formula and Derived Quantities

In some cases, searching also can provide an estimate of the molecular formula as well as two derived values: the nominal molecular weight and the number of rings plus double bonds. To accomplish this, W_s values are first computed for each formula, each molecular weight, and each ring plus double bond value found among the library retrievals. For each of these three properties, the result with the highest W_s value was then selected and marked as being correct or incorrect. From these results, the relationship between these highest W_s values and the probability that each of these properties was correct was derived.

Neutral Losses

Searches were done with and without the use of neutral loss peaks (mass-to-charge ratio measured relative to the molecular ion). Neutral loss peaks between the molecular ion and one-half of the molecular ion were employed in match factor calculations. For match factor computation, these peaks are simply added to the conventional peaks. Tests also were done by using only neutral loss peaks, but results were uniformly less accurate.

Substructure Present Probabilities

For convenience, W_s values, which range from 0 to 1, were divided into 26 segments, and each segment was given a sequence number, w, from 0 to 25. The first segment, $w = 0$, corresponds to $W_s = 0$, whereas the others were equally divided among nonzero W_s values, each spanning a 0.04 range. The number of test compounds that contain substructure s that produce W_s values in segment w is represented as $N_s^+(w)$. The corresponding number of test compounds that do not contain the substructure is $N_s^-(w)$. For any w value, these results can be converted to conventional recall, RC (percent of test compounds containing substructure s that have been identified as such), and percent correct, %C (or reliability [24]), pairs, as follows:

$$\text{RC}(w) = \frac{100\,\Sigma_{i=w}^{25} N_s^+(i)}{\Sigma_{i=0}^{25} N_s^+(i)} \quad (1)$$

$$\%\text{C}(w) = \frac{100\,\Sigma_{i=w}^{25} N_s^+(i)}{\Sigma_{i=w}^{25} N_s^+(i) + \Sigma_{i=w}^{25} N_s^-(i)} \quad (2)$$

While these averaged (integral) measures are appropriate for comparison of and documentation of performance, differential probabilities are better expressions of the results of a single search. These can be expressed in relative or absolute terms:

$$R^+(w) = \frac{N_s^+(w)}{N_s^-(w)} \quad (3)$$

$$P^+(w) = \frac{N_s^+(w)}{N_s^+(w) + N_s^-(w)} = \frac{1}{1 + R^+(w)^{-1}} \quad (4)$$

where $R^+(w)$ and $P^+(w)$ are relative and absolute probabilities that a substructure is present at an observed w value. These probabilities can show significant statistical variations. Therefore, in actual probability calculations, least-squares fits to $N_s^+(w)$ and $N_s^-(w)$ are used in place of their original values.

Substructure Absent Probabilities

A small W_s value will generally produce a small $P^+(w)$ value, which reflects a low probability that the substructure is present in the unknown compound. In this case, it is convenient to express probabilities in terms of substructure absence, $P^- = 1 - P^+$, or $R^- = 1/R^+$. Corresponding RC and %C values may be derived from eqs 1 and 2, respectively, by interchanging + and − superscripts and summing from $i = 0$ to $i = w$. Of course, in this context recall refers to those compounds that do not contain the specified substructure.

A problem, however, remains with reporting substructure absent information. Uncommon substructures will often show high probabilities of being absent that simply reflect their rarity. If, for instance, 5% of

the compounds in the library contain fluorine, there is an implicit 95% probability, before searching, that fluorine is absent. Therefore, a 95% predicted probability of fluorine being absent would simply imply that the search provided no additional information. This issue is discussed in detail by Curry [25]. To avoid this problem, the following corrected relative substructure-absent probabilities are used:

$$R^-_{corr} = R^-(N_{present}/N_{absent})$$

where $N_{present}$ and N_{absent} are, respectively, numbers of compounds in the reference library that contain and do not contain the substructure. This new value reflects the change in confidence, caused by library searching, that the substructure is not in the unknown.

An examination of false negatives (a substructure in the unknown is predicted as being absent), showed that they often arose from searches in which no similar spectra were found. Further examination showed a strong inverse correlation between the highest match factor in a hit list and the likelihood of such a false negative prediction. This is shown in Figure 1 for cases where $W_s = 0$. Incorporation of these trends into probabilities of substructure absence can significantly reduce false positives. For instance, for substructure-absent predictions based on $W_s = 0$, such a correction will decrease the number of erroneous predictions (false negatives) by about two-thirds with only a 15% reduction in recall.

A related, but much weaker, correlation between false *positives* and match factors for the best-matching retrieval also was detected.

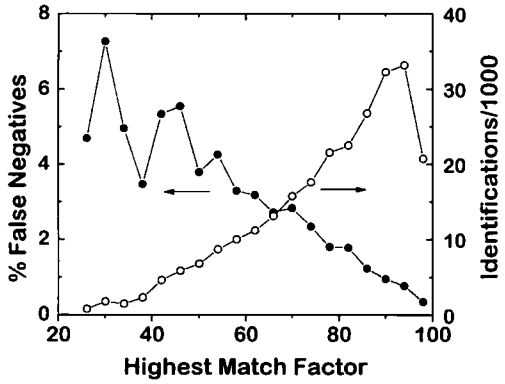

Figure 1. False negatives versus highest match factor when substructure weight is 0. Percent of trials in which a substructure weight (W_s) was equal to zero (no substructures among retrievals), but where the substructure was actually present in the test compound (filled circles). Values are reported at 25 unit intervals and lines are shown only for clarity. Also shown is the corresponding distribution of highest match factors for all substructure-absent identifications (open circles). All substructures in Table 3 were used along with a number of others from refs 12 and 13.

Results

Table 1 shows results for substructures employed in previous comparative STIRS and KNN studies [8]. Comparisons of results of the present system with STIRS are given in columns labeled "With MW." STIRS and these results both use neutral loss peaks and therefore require the molecular weight of the unknown compound as input. Prior KNN studies did not use neutral losses, so neither did the comparisons shown here ("Without MW" columns). All results are given as the percentage of correct identifications at the level of recall given in the earlier studies. By a "correct identification," we mean a search result in which a substructure present in the test compound produces a W_s value above some prespecified minimum. These minimum W_s values are fixed as needed to achieve the desired recall value. A compound in which the substructure is absent, but whose search produces a W_s value above the minimum is a "false positive." The percent of all retrievals having W_s above the specified minimum is, therefore, the percent of correct identifications (%C or reliability [24]). Note that recall pertains only to searches in which substructure s is present in the test compound; it represents the percent of these searches in which substructure s is correctly identified as being present.

Overall results of the present method, shown in columns labeled "Dot full" (dot product algorithm, full library) are dramatically better than these earlier results (columns STIRS and KNN). To find the origin of these differences, searches were done by using a variety of modified algorithms and library sizes (Table 1).

Since the earlier STIRS and KNN studies used a reference library with slightly more than one-quarter of the compounds in the present library, most results given in Table 1 were obtained by using a reduced library that was generated by random rejection of three-quarters of all retrievals. This simulates a library that contains 17% fewer spectra than used in the earlier studies. Also, earlier studies used 500 test spectra, less than one-fifteenth as many as employed in the present study. Significantly larger statistical deviations are therefore expected in the earlier results, especially for the less common substructures.

When the reduced library (Dot, in the "With MW" section of Table 1) is used, identification accuracies of the present algorithm are comparable to STIRS. The slightly higher percent correct value of the present system (85.8% versus 82.8%) is probably not significant given the differences in library and test spectra. The principal source of the improved performance of the present system is clearly the fourfold larger library size. This leads to a threefold reduction in false identifications.

The present system, however, performs far better than the earlier KNN method even when the reduced size test library (Dot, in the "Without MW" section of

Table 1. Percent correct identifications (%C) at fixed recall (RC)

			With MW[a]				Without MW[b]						
				%C				%C					
Substructure[d]	Num. in test set[e]	% in lib.	RC	STIRS[f]	Dot[g]	Dot[c] full	RC	KNN[h]	KE1[i]	KE$\frac{1}{2}$[j]	KDot[k]	Dot[g]	Dot[c] full
C=O	3111	47	31	100	95	99	51	71	78	80	83	88	96
OH	1558	23	42	78	82	95	45	71	62	67	67	74	86
—O—	2348	37	40	98	95	98	54	74	77	81	85	90	96
OCH_2, OCH_3	1806	29	49	81	88	95	36	66	76	80	82	90	97
OC=O	1435	22	55	87	86	94	29	61	81	87	89	92	97
Phenyl(any)	2421	33	75	86	91	96	71	70	84	86	88	93	97
—NH_2	396	5.3	44	69	55	87	17	62	60	65	71	83	92
> NH	430	11	19	58	83	94	25	41	43	55	61	69	75
—N <	875	18	30	89	89	97	51	72	65	65	72	77	87
—S	653	11	35	86	96	100	32	60	83	88	95	97	100
—F	274	6.1	61	84	83	100	18	40	98	99	100	100	100
—Cl	834	8.8	74	96	96	100	54	89	82	91	96	97	100
Alkyl ≥ C_3	1612	19	56	70	79	85	54	68	72	75	77	78	86
Alkyl ≥ C_2	1554	19	19	78	80	95	32	49	55	57	60	65	85
C=C	1202	18	38	88	69	85	42	68	57	60	60	69	83
—CH <	2917	45	44	63	85	92	49	58	77	78	81	82	91
> C <	1464	27	20	85	96	99	41	57	80	80	81	90	96
C-ring	1090	15	50	83	91	98	54	70	85	83	87	89	96
Het-ring	1895	46	31	95	91	98	57	71	68	70	74	82	92
Average			42.8	82.8	85.8	95.1	42.7	64.1	72.7	77.3	79.4	84.5	92.2

[a]Uses neutral loss peaks based on input molecular weight.
[b]Does not use neutral loss peaks, hence does not require input molecular weight.
[c]Present method with complete library.
[d]From ref 8. Except for ring-only substructures, they may be in a ring or chain. All H-atom attachments are explicit. Phenyl(any) is an isolated benzenoid ring; C-ring is a ring that contains only C-atoms; Het-ring is a ring that contains one or more non-C atoms.
[e]Number of compounds that contain the substructure among the 7891 spectra test set.
[f]From ref 8 (500 spectra test set).
[g]Present method with reduced NIST/EPA/NIH library (one-quarter full size).
[h]K-nearest neighbor results from ref 8 that use three-fifths voting scheme, Euclidean distance, first order scaling.
[i]Present implementation of KNN (footnote h) with one-quarter size library.
[j]KNN that uses the square root of Euclidean distance with one-quarter size library.
[k]KNN that uses the dot-product distance measure with one-quarter size library.

Table 1) is used. This is primarily due to several modifications made to the original KNN procedures. Starting from the original KNN method, the cumulative effect of each of these modifications on search performance is shown in Table 1. The original KNN system used a simple Euclidean distance to measure spectral dissimilarity and a simple voting scheme to classify compounds. Our implementation of the same system (KE1) gives similar, though somewhat better results (72.7% versus 64.1%). These differences presumably arise from statistical effects due to the small test used in the earlier studies as well as from the use of different reference libraries. Performances were redetermined after sequentially making the following modifications: (1) scaling abundances by their square roots (KE1/2), (2) replacing the Euclidean distance with the dot product comparison function (KDot), (3) replacing the simple voting scheme with the present distance-based weighting method (Dot), and (4) increasing library size by a factor of 4 (Dot full).

A comparison of the present system with more recent STIRS results is given in Table 2. This STIRS study used a larger library of 25,598 organic compounds (the earlier studies used nonorganics also), 899 test spectra, a newly defined set of substructures, and somewhat modified algorithms. These results are compared to results of the present method for all substructures contained in at least 40 test compounds in the original study. These comparative studies also used only organic compounds. The reduced size library contained one-half of the spectra in the full-sized library, which simulated a library with 15% more spectra than the STIRS studies. By using this reduced library, performances are again comparable.

The overall performance of the present system is documented in Table 3 using the substructures in Tables 1 and 2 and others of general interest. The effectiveness of substructure recognition is presented for searches with and without the use of neutral loss peaks (labeled "With MW" and "Without MW," re-

Table 2. Percent correct identifications[a] (%C) at fixed recall (RC)

Substructure	Num. in test set	% in lib.	RC	%C STIRS	%C Dot[b]	%C Dot full[c]
$-C_6H_5$	974	15	72	87	78	88
$-OSi(CH_3)_3$	448	5.1	[100	**93**	90[d]	94[d]]
$ArOCH_3$	405	7.7	56	86	91	95
$-CO-OCH_3$	390	6.7	70	75	73	88
$-CH_3$	5714	76	51	99	99	99
$Ar-O-$	488	10	56	88	75	82
$Ar-OH$	424	5.8	43	80	76	87
$-OCH_3$	975	18	44	91	97	98
$-CO-O$	1291	20	53	90	94	96
$-CO_2H$	266	3.5	41	70	68	88
$-OCH_2-$	931	13	42	84	85	90
$-Si(CH_3)_3$	491	6.0	[100	**67**	99[d]	99[d]]
ArCl	474	4.4	58	79	95	96
$-CO-O-CH_2$	469	6.6	57	71	78	87
.CH.CH.O	195	5.0	64	82	65	77
$-(CH_2)_3-$[e]	1356	15	52	75	89	91
$-CH(CH_3)_2$[e]	588	6.8	27	48	50	58
$-CO-CH_2-CH_2$[e]	493	5.6	51	54	86	91
$-CO-NH$[e]	233	5.1	29	67	81	93
Totals[f]			50.9	78.0	81.2	88.5

[a]Except where noted, substructures and STIRS results are from ref 12 (899 spectra test set, 25,598 library compounds). Dashes (—) represent chin (nonring) bonds; periods (.) represent ring bonds.
[b]Present system used with the reduced NIST/EPA/NIH library (one-half full size).
[c]Present system with full NIST/EPA/NIH library.
[d]Recall values computed at percent correct are shown in boldface [12]. 100% recall in ref 12 could not be achieved by the present method.
[e]STIRS results from ref 18b.
[f]Does not include values in brackets.

spectively). Recall values are given for two accuracy requirements. One is an average reliability (percent correct) of 90%. The other is for a probability of being correct, P^+, of at least 90%. The former values are useful for comparison of search performances of different systems, whereas the latter has a clearer practical meaning.

The ability of the present method to predict the absence of a substructure is given in two ways. First, the reliability of a substructure-absent prediction is given along with its recall value for searches where the substructure is not found in the hit list ($W_s = 0$). Also, recall values for substructure-absent probabilities, P^-, of at least 90% are given. Note that these values are corrected for prior occurrence probabilities as described in the foregoing text. However, the corrections discussed earlier for false positives were not made.

Also shown in Table 3 are results for identification of chemical formulas, molecular weights, and rings plus double bonds. Results also are shown for searches in which peak abundances were multiplied by the square of their mass-to-charge ratio values.

Discussion

The ability of a library search system to identify a substructure depends on two factors: (1) the number and diversity of library compounds that contain the substructure and (2) the strength of a substructure's "signature" and the ability of the system to perceive it.

An increased representation in the library will increase the number of chemical environments that the substructure is "embedded" in and therefore will increase the likelihood of finding similarly configured substructures in an unknown compound. Identification of even common functional groups can benefit from an increased representation. This is evident in Tables 1 and 2, which show that identification of even very common groups, such as carbonyl, benefits from increased library size.

The strength and uniqueness of a signature and the ability of a search method to perceive it is of course crucial for substructure identification. A sufficiently unique signature can fully compensate for a small representation in the library. The trimethylsilyl group,

Table 3. Substructures identified (recall, RC) at fixed levels of accuracy

Substructure[c]	No. in test set	% in lib.	Feature present[a] RC values: Without MW %C = 90	Without MW $P^+ > 0.9$	With MW %C = 90	With MW $P^+ > 0.9$	Feature absent[b] Without MW $W_s = 0$ RC	%C	$P^- > 0.9$
Elements and compound class									
F	274	6.1	73	64	75	61	78	92.5	
Cl	834	8.8	82	66	89	76	66	94.0	86
Br	317	3.1	72	56	83	56	76	94.2	
I	60	0.70	58	45	65	63	93	82.1	
N	2484	44	88	70	90	69	37	98.7	77
O	5156	76	96	73	98	75	16	99.9	67
S	653	11	70	58	72	56	62	94.6	84
P	156	2.5	70	53	72	58	91	94.4	97
HC	1053	6.1	88	75	95	76	61	99.7	87
Sat'd. HC	320	1.6	74	58	77	61	83	99.6	85
Unsat'd. HC	733	4.5	80	53	89	72	67	99.2	85
Aromatic	3559	50	98	86	99	84	71	99.3	88
CH & O	2571	28	85	61	91	70	27	100.	76
Substructures from refs 11 and 12									
Alkyl C_3+	3002	39.9	79	58	79	56	18	98.9	60
C=O	3111	46.9	70	44	76	51	15	98.5	48
OH	1558	22.6	40	23	58	40	19	99.0	43
OC=O	1435	22.0	36	55	47	67	30	96.6	53
Phenyl(any)	2421	32.5	88	76	90	72	63	99.1	80
$-NH_2$	396	5.3	25	20	40	21	61	92.6	
> NH	430	11.1	28	23	28	18	58	91.1	
> N—	875	17.7	44	30	44	29	57	95.8	76
OCH_2, OCH_3	1806	29.0	65	45	67	45	27	98.7	62
C-ring	1090	15.5	69	56	75	51	47	97.1	73
$-Si(CH_3)_3$	591	6.0	97	92	99	99	90	99.3	97
Cyclohex	209	2.3	51	42	57	39	85	90.9	
$PhCO_2H$	58	0.64	16	16	18	18	92	92.5	
C=C	1202	18	17	12	19	15	18	97.7	45
—O—	2348	37	74	53	80	59	21	99.2	65
—CH <	2917	45	53	37	48	37	4	100.	36
> C <	1464	27	58	45	61	47	23	96.1	50
Het-ring	1895	46	62	45	66	40	28	98.7	64
$-C_6H_5$	974	15.2	68	56	67	50	64	98.7	79
$-OSi(CH_3)_3$	448	5.1	96	70	98	76	92	99.5	97
$ArOCH_3$	405	7.7	66	45	73	45	76	98.0	80
$CO-OCH_3$	390	6.7	56	40	68	56	67	92.7	
$-CH_3$	5714	76.1	95	74	97	76	5	99.7	54
Ar—O—	488	10.3	7	5	33	33	70	97.7	75
Ar—OH	424	5.8	23	23	28	22	70	96.3	75
$-OCH_3$	975	18.3	49	48	74	67	45	97.9	67
CO—O	1291	19.5	59	45	73	58	35	97.6	59
$-CO_2H$	266	3.5	20	30	38	20	61	87.1	71
$-OCH_2-$	931	13.4	43	27	47	19	39	97.6	63
$Si(CH_3)_3$	491	6.0	97	92	98	88	90	99.3	97
ArCl	474	4.4	79	65	80	65	89	95.7	
$CO-O-CH_2$	469	6.6	37	27	51	40	61	94.4	

Table 3. (continued)

Substructure[c]	No. in test set	% in lib.	Feature present[a] RC values: Without MW %C = 90	Without MW $P^+ > 0.9$	With MW %C = 90	With MW $P^+ > 0.9$	Feature absent[b] Without MW $W_s = 0$ RC	%C	$P^- > 0.9$
.CH.CH.O	195	5.0	29	20	37	24	78	90.1	
$-(CH_2)_3-$	1356	14.8	49	35	54	35	37	97.4	61
$CH(CH_3)_2$	588	6.8	9	7	7	6	44	94.4	53
$CO-CH_2-CH_2$	493	5.6	53	29	53	38	65	93.5	84
—CO—NH	233	5.1	24	21	33	27	73	87.	
Formula and related									
Formula	18	28							
	34[d]	21[d]							
MW	39	23							
	54[d]	34[d]							
Rings + double bonds[e]	60	38							
	58[d]	34[d]							

[a]Substructure-present identifications that use neutral losses (With MW) and without neutral losses (Without MW). Recall values are given for two requirements: (1) 90% of all identifications are correct (%C = 90); (2) the probability of each identification being correct is greater than 0.9 ($P^+ > 0.9$).

[b]Substructure-absent identifications. No neutral loss peaks are used (Without MW). Recall and percent correct values are given weight factor $W_s = 0$ (no substructures in hit list). Also shown are percent of correct substructure-absent predictions where the probability of being correct is greater than 0.9. All values are corrected for substructure-occurrence probability as discussed in the text and do not use corrections for top match factor (Figure 1).

[c]Nonring bonds are denoted by lines (—, >, =); ring bonds are denoted by periods (.). HC = hydrocarbon, CH & O = contains C, H, and O atoms and no others, Ar = aromatic atom, Ph = single benzenoid ring.

[d]Peaks are multiplied by the square of their mass-to-charge ratio values for match factor determination.

[e]Number of rings and double bonds, also known as double bond equivalents.

for instance, is the best identified of all substructures even though it is present in only 6% of library compounds. On the other hand, certain structural features that have no unique fingerprint can often be identified when present as a part of larger substructures that do have characteristic fingerprints. For instance, although tertiary carbon atoms (—CH <) themselves have no characteristic fingerprint, their presence can be identified with good reliability because a large proportion of structural groups that contain tertiary carbon atoms do have clear signatures.

Neutral Losses

Certain substructures tend to be readily eliminated from ionized molecules. Their expulsion leads to the formation of "primary neutral loss" peaks at mass-to-charge ratio values lower than the molecular ion by an amount equal to their mass. For these substructures, identification can be improved by using neutral loss peaks in addition to conventional peaks for match factor determination. This improvement is evident for several of the substructures in Table 3, such as hydroxyl, carboxyl, amino, and ester groups. The practical problem with this approach is the requirement that the molecular weight be known in advance. Conventional electron ionization mass spectrometry cannot be reliably provided this value. Approximately 20% of library compounds show no easily identifiable molecular ion.

Reference Library

As is evident from the major improvements in identification accuracy gained by increased library size (compare Dot and Dot full columns in Tables 1 and 2), to obtain the most reliable results, a large, comprehensive, structure-based library is required. Furthermore, it is not clear that such libraries can ever be too large. Any increases in the number of compounds represented by good quality spectra appear certain to further improve the ability to identify substructures, even if the compounds added are themselves not of direct interest.

Performance

Spectrum screening retrieves an average of 450 spectra per search, or about 0.7% of all library spectra, for subsequent peak-by-peak comparison to the submitted

unknown spectrum. As implemented, the comparison step consumes about 80% of the total search time. Screening, therefore, reduces overall search times by over a factor of 100, which results in search speeds that are comparable to those of a conventional compound identification search. If desired, tighter screening could further reduce search times with little loss in identification accuracy. Tests showed that a twofold reduction is the average number of spectra that passed through the screen had a barely measurable effect on performance (1–2% reduction in recall).

Peak Weighting Schemes

Identification of substructures that have characteristic peaks might be expected to improve if these peaks were specifically weighted. A number of such schemes were tested, but, as described in the following text, none of them markedly improved substructure identification accuracy.

Because low mass peaks are often more important for substructure identification than high mass peaks [23], attempts were made to improve performance by increasing the relative weighting of low mass peaks. This is the opposite of what is typically done for compound identification [3], where the highest mass peaks in a spectrum are the most diagnostic. Several schemes that used different weighting functions and mass ranges were applied, but none of them improved overall performance. For some substructures, modest improvements of 2–3% recall at a fixed reliability were observed, but reductions in performance were observed for others.

A more specific peak weighting scheme was tested for peaks that belong to the "aromatic series" [11]. Although this resulted in a noticeable improvement for identification of aromatic substructures, effects were small. For instance, at 90% percent correct for singly substituted phenyl, recall increased from 68 to 70% (neutral loss peaks were not used). For chloroaromatics (ArCl), corresponding recall values increased from 79 to 82%.

A number of other substructure-specific weightings schemes were implemented, with similar results. The present unweighted match factor appears to be near optimal for general purpose use. Only modest improvements appear possible with the use of substructure-specific weighting schemes. It is not clear whether these modest improvements justify the added complexity and risk of "overtraining."

For the case of molecular weight estimation, however, where peaks near the molecular ion contain important information, increasing the contribution of high mass peaks to the match factor significantly improved performance. As shown in Table 3, after multiplying peaks by the square of their mass-to-charge ratio values, a nearly 50% improvement in recall was observed. A modest improvement in formula prediction resulted from this scaling, but ring plus double bond prediction actually worsened.

Probability Estimation

The mass spectral comparison function used here for substructure identification provides a single overall measure of spectral similarity and, except for the absence of peak weighting, is identical to that commonly used for compound identification. It relies on the simple premise that the more similar two spectra are, the more likely it is that the compounds that produced them have substructures in common. This obviously applies to the limiting case of nearly identical spectra for a single compound. As pairs of spectra become more dissimilar, the likelihood that the compounds that produce them have substructures in common diminishes. According to the present similarity measure, a difference of 75 match factor units implies a reduction in this likelihood by a factor of 2.

The sum of the weights of all retrievals (the hit list normalization factor, N, defined earlier) may be viewed as the effective number of reference spectra used for deriving probabilities. Its average value was 6.7, which indicates that the top hit, on average, contributed about one-seventh of all of the substructure information.

The derived probability of a substructure being in the unknown compound, P^+, is related to the probability, described in earlier work [3], that a retrieved compound precisely matches the unknown compound. Both were derived from relative match factors that, in turn, correlated with the probability that a retrieval was correct. For probability estimation, however, the earlier work made direct use of the observation that differences in match factors were directly related to relative probabilities that a retrieval matched the unknown compound. This direct approach could not be used here because for substructure identification, multiple retrievals can be correct (i.e., contain a substructure present in the unknown compound). A simple exponential function, discovered by trial and error, was used instead. Also, probabilities of matching the unknown compound reported in the earlier work were roughly four times as sensitive to match factors than are the substructure-present probabilities.

Earlier studies [26] showed that for exact compound matching, the highest match factor in the hit list was related to the probability that the compound was in the library. No equivalent relation could be derived for the present analysis. However, in the present work, the highest match factor value did (inversely) correlate with the likelihood of falsely predicting that a substructure was absent in the unknown compound (Figure 1).

Comparison to STIRS

Tables 1 and 2 show that STIRS and the present method give comparable results when used with reference

libraries of comparable size. Because of the different test sets and reference libraries in the two studies, it is not possible to usefully discuss differences in detail.

The present distance-based retrieval weighting scheme is quite different from the equal weighting assumption for the top 15 retrievals in STIRS. The effective number of retrievals used by the present scheme depends on the match factors found; the number ranges from near unity when only one retrieval is a good match to up to 25 when all retrieved spectra are equally similar to the spectrum of the test compound.

A screening strategy has not been proposed for STIRS. Two other key differences are the requirement by STIRS that the molecular weight of the unknown be provided and the inability of STIRS to report substructure-absent probabilities. Also, the large test set in the present work allows easier to interpret differential probabilities (P^+ and P^-) to the reported. The reported lack of sensitivity of STIRS performance to library size [13] is a surprising difference.

The use of multiple hit lists, each created from a different set of mass spectral features, has been shown to significantly benefit the performance of STIRS [18]. An examination of the benefits of the addition of such features to the present system is underway.

Comparison to the K-Nearest Neighbor Method

Isenhour and co-workers [8] have reported a method that identifies substructures in an unknown compound by using the K-nearest library spectra as measured by their Euclidean distance from the unknown spectrum. A substructure is considered to be present in the unknown when it is contained in a prespecified number of the K-nearest library compounds. This method did not perform as well as STIRS in a comparative study [8].

The present system may be viewed as an improved version of this KNN procedure. Effects of each improvement are shown in Table 1 for a three out of five "voting" scheme (identification requires that three or more of the five "nearest" library compounds contain the substructure). The percent correct at fixed recall for the present implementation of the original algorithm are shown in column KNN in Table 1. They show trends similar to the original implementation, but overall results are 8.6% better. Some of this difference may be statistical because only 500 test compounds were used in the earlier studies. The biggest single difference in performance is for fluorine, which was present in only 16 of the original test compounds. Exclusion of this one value reduces the differences to about 6%.

Column KE$\frac{1}{2}$ in Table 1 shows an improvement of 4.6% correct that results from square-root scaling of the abundance. Replacement of the Euclidean distance with the dot product function results in a further 2.1% improvement (KDot). Previous compound identification studies [26] showed similar performance gains for such modifications.

The next column, Dot, shows a further 5.1% improvement when the present retrieval weighting scheme replaced the original three out of five voting. Finally, a fourfold increase in library size (Dot full) showed the largest increase in percent correct, 7.7%, which corresponds to a 50% decline in missed identifications. The overall improvement in reliability due to algorithm improvement and library size is dramatic; it goes from 72.7 to 91.2% correct at fixed recall, or a threefold decline in incorrect identifications.

Comparison to MSNet

An artificial neural network method for identification of substructures from mass spectra has been reported by Curry and Rumelhart [19]. It employed a large number of mass spectral features, most of them taken from STIRS, along with a reference library of 31,926 library spectra. By using a test set of 12,671 spectra, results were found to be similar, though somewhat better than results reported for STIRS in ref 12. Considering the significantly smaller library size in the STIRS studies, the inherent ability of the two systems to identify substructures appears to be similar. This suggests, in turn, that the ability of MSNet to identify substructures is comparable to the present system.

Three disadvantages that were cited for library-search systems by Curry and Rumelhart [19] and repeated by Warr [5], include: (1) lack of absolute identification probabilities, (2) inability to report substructure-absent probabilities, and (3) slow search speed. All three deficiencies are eliminated in the present system.

Conclusions

A practical library search procedure that extracts chemical substructure information from conventional electron-ionization mass spectra has been developed and tested. By using results of a library search, this method derives probabilities that a given substructure is present or absent in the unknown compound. The reliable reporting of substructure-occurrence probabilities was significantly enhanced by selection of optimal methods for processing spectral match factors as well as by using a large test set. Advance knowledge of the molecular weight of the unknown compound is not required, although identification of certain substructures can benefit if it is provided. Identification of all substructures, even very common ones, has been shown to benefit greatly from an increase in library size. An efficient spectrum screening procedure has been designed to enhance the speed of substructure identification, which resulted in search times comparable to those typical of conventional compound identification searches. The overall performance, when linked

to a large structure-based mass spectral library, is sufficient to recommend it for routine use as a first step in the structural elucidation of compounds not represented in reference libraries.

References

1. Warr, W. A. *Anal. Chem.* **1993,** *65,* 1045a–1050a.
2. Martinsen, D. P.; Song, B.-H. *Mass Spectrom. Rev.* **1985,** *4,* 461–490.
3. Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 859–866.
4. Zurcher, M.; Clerc, J. T.; Farkas, M.; Pretsch, E. *Anal. Chim. Acta* **1988,** *206,* 161–172.
5. Warr, W. A. *Anal. Chem.* **1993,** *65,* 1087a–1095a.
6. Kowalski, B. R.; Bender, C. F. *Anal. Chem.* **1972,** *44,* 1405–1411.
7. Justice, J. B.; Isenhour, T. L. *Anal. Chem.* **1974,** *46,* 223–226.
8. Lowry, S. R.; Isenhour, T. L.; Justice, J. B., Jr.; McLafferty, F. W.; Dayringer, H. E.; Venkataraghavan, R. J. *Anal. Chem.* **1977,** *49,* 1720–1722.
9. Damen, H.; Henneberg, D.; Weimann, B. *Anal. Chim. Acta* **1978,** *103,* 289–302.
10. Domokos, L.; Henneberg, D.; Weimann, B. *Anal. Chim. Acta* **1984,** *165,* 61–74.
11. Kwok, K.-S.; Venkataraghaven, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973,** *95,* 4185–4194.
12. Dayringer, H. E.; Pesyna, G. M.; Venkataraghavan, R.; McLafferty, F. W. *Org. Mass Spectrom.* **1976,** *11,* 529–542.
13. Haraki, K. S.; Venkataraghavan, R.; McLafferty, F. W. *Anal. Chem.* **1981,** *53,* 386–392.
14. Henneberg, D.; Weimann, B.; Zalfen U. Poster Presentation at the 12th International Conference on Mass Spectrometry, Amsterdam, **1991.**
15. (a) Munk, M. E.; Christie, B. D. *Anal. Chim. Acta* **1989,** *216,* 57–78; (b) Christie, B. D.; Munk, M. E. *J. Amer. Chem. Soc.* **1991,** *113,* 3750–3757.
16. Funatsu, K.; Miyabayashi, N.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 18–28.
17. Gasteiger, J.; Hanebeck, W.; Schulz, K.-P. *J. Chem. Inf. Comput. Sci.* **1992,** *32,* 264–271.
18. (a) Dayringer, H. E. Ph.D. Thesis, Cornell University, **1976;** (b) Haraki, K. S. Ph.D. Thesis, Cornell University, **1980.**
19. Curry, B.; Rumelhart, D. E. *Tetrahedron Comput. Methodol.* **1990,** *3,* 213–237.
20. Werther, W.; Lohninger, H.; Stancl, F.; Varmuza, K. *Chemom. Intell. Lab. Syst.* **1994,** *22,* 63–76.
21. NIST/EPA/NIH Mass Spectral Database with Selected Replicate Spectra; Standard Reference Database No. 1A; NIST: Gaithersburg, MD, **1992** (PC Version 4.5).
22. Sokolow, S.; Karnofsky, J.; Gustafson, P. The Finnigan Library Search Program; Finnigan Application Report 2; March, **1978.**
23. Dayringer, H. E.; McLafferty, F. W. *Org. Mass Spectrom.* **1976,** *11,* 543–551.
24. McLafferty, F. W. *Anal. Chem.* **1977,** *49,* 1441–1443.
25. Curry, B. in *Computer-Enhanced Analytical Spectroscopy;* Plenum: New York, **1990;** pp 183–209.
26. Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 316–323.