The Evaluation of Mass Spectral Search Algorithms

G. T. RASMUSSEN and T. L. ISENHOUR*

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

Received February 20, 1979

Mass spectral search algorithms are tested by two methods. Trial searches of target spectra are used to compare the effects of various data-encoding methods and different distance metrics on search results. The effectiveness of selected search algorithms in retrieving similar compounds is assessed through the use of a propagation method, which employs repetitive searches of the nearest matches to a "seed" compound. A library containing the spectra of nearly 17 000 organic compounds is used with these methods to provide an evaluation of the relative performance of different search techniques.

The computerized search of mass spectral data for the identification of organic compounds has become an important tool in analytical chemistry. Three factors that have prompted an increased interest in computer searches are (1) the widespread availability of data acquisition systems that produce digitized mass spectra, (2) the vast amount of data that can be generated in a short time by using such systems to monitor the effluent of chromatographic columns, and (3) the growth of computer-compatible mass spectral data files to the extent that such files routinely contain tens of thousands of spectra. An effective mass spectral search algorithm can aid the analyst in two ways. In many routine cases, a search can correctly identify the unknown spectra which have corresponding library entries. When the unknown is not contained in the reference file, a search can provide useful chemical information about the unknown by listing compounds with similar spectra. Although a wide variety of different search methods have been described, only limited efforts have been made to compare these techniques. In this paper, mass spectral searches are evaluated and compared by two methods. One method, which relies on trial searches of known spectra, is a common technique for testing search performance. This method is useful in assessing the ability of a search system to identify specific spectra. The second method, which offers a new approach to the problem, is based on repeated searches of the nearest matches selected by a search beginning with a "seed" spectrum. This method gives information about the ability of a search to identify the spectra of chemically similar compounds when the unknown spectrum is not contained in the library file.

In a conventional library search system, the unknown spectrum is compared with the spectra of known compounds which are contained in a reference collection called the library. The two basic elements of a computerized search are the data-encoding method and the comparison algorithm. The data-encoding method determines what information from the original mass spectra is actually available to be used in the search. Generally, mass spectra can be treated as points or vectors in a multidimensional space, with each mass position representing a separate dimension and each intensity value indicating the length of the vector as measured along the axis representing the corresponding mass position. The comparison algorithm defines the manner in which unknown and reference spectra are compared. The distance metric used by the comparison algorithm provides a quantitative measure of the distance between vectors representing the reference and unknown spectra, and this distance reflects the dissimilarity between spectra. Prefilters, which are simple preliminary comparisons between reference and unknown spectra to determine whether a detailed distance computation is necessary, are sometimes included in a comparison algorithm. By eliminating from consideration many reference spectra which are not likely to be similar to the unknown, prefilters can save the computation of distances for many reference-unknown pairs during a search. Prefilters can also affect search performance by ruling out reference spectra which are similar to the unknown but differ from it in some specific and undesirable way.

Before computer searches were developed, manual searches of low resolution mass spectral data encoded on punched cards were used.[1,2] The first published suggestion of a computerized mass spectral search has been attributed to Abrahamsson, Stallberg-Stenhagen, and Stenhagen.[3] In the years following the publication of this brief note, a variety of data-encoding methods and comparison algorithms were investigated.[4-9] In general when a search system is described in the literature, the evaluation of its performance is reported by listing the results of trial searches. Typically these trial searches are run on spectra of target compounds, which are known to have corresponding library entries. In order to provide a true test of the search system, care is taken to ensure that the target spectra are from a source that is independent of the library file. However, because different target spectra and different libraries are used by different researchers, direct comparison of the results for different search methods are rare. Using a library containing 6880 spectra and a set of 125 target spectra, Grotch has examined the effects of various prefilters, data-encoding methods, and distance metrics on search performance.[10,11] Mathews and Morrison have evaluated a wide variety of search methods but confined their attention to the identification of terpenes, using a library of 122 mass spectra.[12] A later study by Mathews has compared different methods of identifying alkylbenzene compounds from mass spectral data, including library search techniques, with a library of 105 spectra.[13] Although this latter approach which uses a restricted library file may be effective for testing the ability of a search algorithm to distinguish between the spectra of closely related compounds, a complete evaluation of search performance should include trial searches using a large, general library and target spectra of diverse compounds. The portion of this paper dealing with trial searches compares the effects of employing different data encoding methods and distance metrics by performing searches with a set of 40 target spectra and a library containing nearly 17 000 entries. In addition to some of the search strategies previously compared, this study includes the data compression method described by Wangen et al.,[14] the use of selected mass positions for binary spectra as reported by van Marlen and Dijkstra,[15] and a data-encoding method based on an ion series approach to data compression.[16] All techniques considered here are "forward" search techniques which rely solely on mass position and intensity information contained in the mass spectra. Search methods not evaluated

0095-2338/79/1619-0179$01.00/0

in this comparison include interactive searches, reverse searches, text searches, and searches incorporating heuristic data classifications.[17-21]

To complement the comparative search evaluation based on trial searches, this paper introduces a propagation technique which is intended to examine the ability of a search algorithm to provide meaningful chemical information about the compound producing the unknown spectrum. In applying the propagation technique, the mass spectrum of a "seed" compound drawn from the library file is searched as an unknown spectrum would be. The spectra of the n nearest matches found by the search algorithm are then searched, and in turn the spectra of the n nearest matches to each of these spectra. As the process continues, a list of the compounds found by the repeated searches is kept. Because a search algorithm will correctly identify a spectrum drawn directly from the library file, a maximum of $n - 1$ new compounds can be introduced with each search. For example if n equals 3, the three nearest matches to the seed spectrum will be the seed spectrum itself and the spectra of two other compounds. The spectra of these two compounds are searched and their three nearest matches are recorded. As the subset of compounds on this list grows with repeated searches, one can observe how the cluster defined by the search algorithm spreads out, or propagates, in the library file. As a propagation continues, more and more compounds may be included until, after many steps, all entries in the library are drawn into the subset. This occurs unless the n nearest matches to each member of a subset are themselves subset members. In this case the subset is considered "closed" because subsequent searches will introduce no new compounds. After several steps of such a propagation, an examination of the chemical and structural characteristics of the compounds in the subset can identify which molecular feature(s) of the seed compound were recognized by the search algorithm under consideration. The homogeneity or diversity of compounds in the propagation subset will reflect the ability of the search algorithm to recognize chemical information. Search algorithms which form closed subsets of similar compounds are considered superior to those which include unrelated compounds in growing subsets. Besides examining the molecular features of compounds in a subset, one can compare the rates at which subsets grow for propagations with the same seed compound but different search algorithms. This rate of propagation can be considered another indicator of search performance. One significant advantage of the propagation technique is that its applicability is independent of the presence or absence of particular compounds in the library file. The method can be used with large or small library files, and seed compounds can be selected either randomly or because of an analyst's specific interest in them. The propagation method will be considered here in some detail, and several search algorithms will be compared by an examination of results for selected seed compounds.

EXPERIMENTAL SUMMARY

The reference library used consisted of the mass spectra of 16 924 unique compounds drawn from the Registry of Mass Spectral Data.[22] Spectra of compounds containing only the elements carbon, hydrogen (including deuterium), oxygen, nitrogen, sulfur, phosphorus, and the halogens were retained in this data file. All mass spectral peaks at integral mass positions with intensities above a 1% threshold relative to the base peak were included to produce a file with data for more than 816 000 peaks at mass positions up to nearly 660 amu. Separate data files were generated for each data-encoding method considered, and a single data file containing the names, molecular weights, and the molecular formulas of all compounds was compiled. Target mass spectra for trial searches were taken from the problems presented in the last three chapters of the text by Silverstein and Bassler.[23] Of these, 40 spectra had corresponding library entries. The target spectra were digitized by hand and recorded on punched cards. Data files containing lists of nearest matches to compounds selected in propagation experiments were generated for convenience in processing this information. All data sets were stored on a magnetic disk package, and all programs were written in Fortran IV and run on IBM 360/75 or 370/155 computers operating at the University of North Carolina Computation Center.

SEARCHES USING DETAILED INTENSITY INFORMATION

In evaluating different search algorithms, one goal is to find a method of encoding and comparing mass spectra that will correctly identify unknown spectra. The unknown spectrum will generally be similar but not identical with the spectrum of the same compound stored in the library file. The mass spectral information available consists of mass position and relative intensity information, and searches use some or all of this information in comparing spectra. The first searches considered are those which use detailed intensity information recorded to a resolution of 1% relative to a most intense (base) peak. The effects of employing various peak selection methods, different distance metrics, and different normalization techniques are examined. Mass spectral peak intensities are usually reported relative to a base peak, but other normalizations have been suggested.[6] Two alternative methods considered here are the total ion current normalization and the normalization of mass spectral vectors to unit length. The conditions of these normalizations are indicated in eq 1 and 2, respectively, where I_j symbolizes the intensity of the peak at the jth mass position and n is the number of mass positions.

$$\sum_{j=1}^{n} I_j = 1 \tag{1}$$

$$\sum_{j=1}^{n} I_j^2 = 1 \tag{2}$$

Three distance metrics for the comparison of mass spectra with detailed intensity information are also considered. These are the summed absolute values of intensity differences, the Euclidean distance, and the similarity index developed by Hertz, Hites, and Biemann.[6,9,10] Equations 3 and 4 detail the first two of these respectively, where D represents the distance, U and L denote the unknown spectrum and the library entry, and other symbols have their previous meanings. The more

$$D = \sum_{j=1}^{n} \text{absolute value } (I_{Uj} - I_{Lj}) \tag{3}$$

$$D = [\sum_{j=1}^{n} (I_{Uj} - I_{Lj})^2]^{1/2} \tag{4}$$

complex Biemann similarity index is based on the average weighted ratios of unknown and reference peak intensities and on the fraction of unmatched intensities in the two spectra. The various peak selection methods considered are taken to be representative of the many different approaches tried in the past. Results for searches using information for all peaks in a spectrum are compared with those for searches relying on a fixed number of peaks selected from the entire spectrum and for searches employing only peaks selected from 14 amu intervals across the spectrum. In these comparisons, results are reported with a graph of the number of target compounds correctly identified as a function of the number of nearest matches that must be considered to include the correct compound. This method is similar to that used by Grotch.[10]

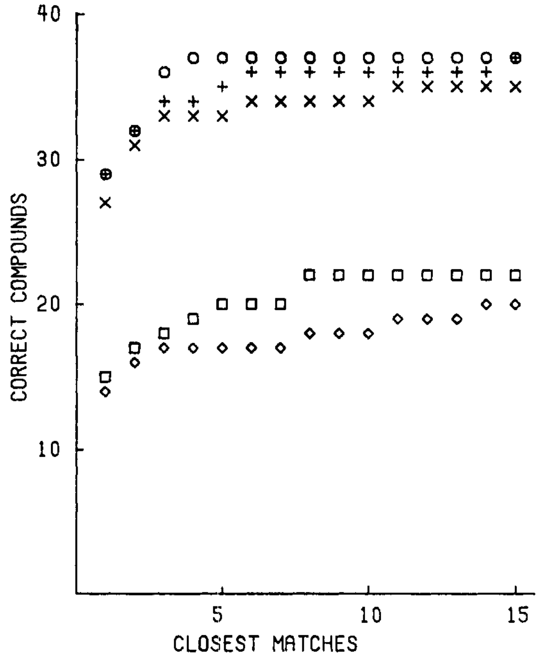


Figure 1. Results for searches with different peak selection methods: ○, complete spectra; ◇, ten most intense peaks; □, ten most significant peaks; +, two peaks per 14 amu; ×, one peak per 14 amu.

However, in these comparisons, isomeric compounds are not counted as correct identifications. When a series of isomeric compounds was found by a search, the correct compound was only rarely displaced by two or more of its close isomers so that the effect of this requirement was minor.

The results obtained by using base peak normalized spectra and an absolute value distance metric with five different peak selection methods are shown in Figure 1. When retaining all peaks, the average spectrum in the library has about 50 peaks. One peak selection method is to retain only the n most intense peaks in a spectrum, and this approach is represented by the case where n equals 10. In order to give more weight to peaks at higher mass positions, peaks have also been selected on the basis of "significance", which is defined as the product of a peak's intensity and mass position.[4,24] Although the results for the search using "significance" rather than intensity as the criterion for the selection of best peaks are somewhat better, only about half of the target compounds are identified within their 15 nearest matches in either case. Another approach to peak selection is to keep the n most intense peaks in 14-amu segments of the spectrum. If n equals 2, the average spectrum in the library file has about 20 peaks while, if n equals 1, the average number of peaks per spectrum is slightly greater than 11. The former case represents the data-encoding method recommended by Hertz et al.,[9] and the latter one reflects an approach studied by Grotch with spectra having only one or two bits of intensity resolution.[11] Although the performance for the search using information for all peaks is best, the results for the peak selection methods which retain n peaks in 14-amu intervals both come quite close.

Results for the searches using different distance metrics are summarized in Figure 2. Base peak normalized spectra encoded with the two most intense peaks in 14-amu intervals were used in these tests. The prefilters normally employed with the Biemann search algorithm were disabled to provide a true test of the distance metrics. The elimination of these prefilters had an insignificant effect on the search performance, but execution time for the search was increased dramatically. In normal operation the prefilters screen out over 99% of the reference spectra so that a similarity index computation is performed for less than 1% of the library entries on the average. The results suggest that the three metrics are about equally suitable for the comparison of mass spectra.

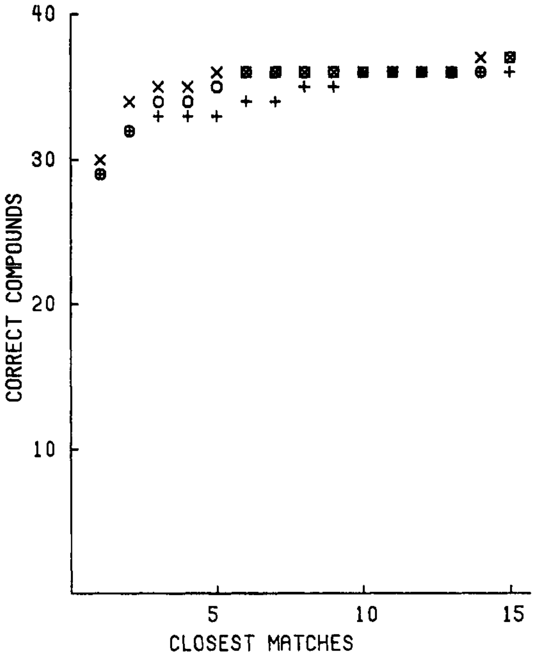


Figure 2. Results for searches with different distance metrics: ○, absolute value distance; +, Euclidean distance; ×, Biemann metric.

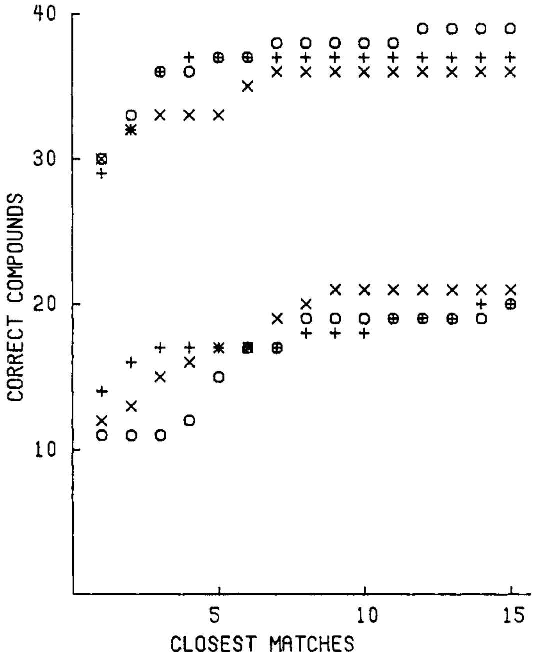


Figure 3. Results for searches with different normalizations: +, base peak normalized; ○, total ion current normalized; ×, unit vector length normalized.

Figure 3 shows the effects of different normalization methods on search performance. Results are reported for three normalizations with two methods of peak selection. For each normalization, the upper row of symbols indicates the results for searches which use complete spectra, and the lower row reflects the results observed using only the ten most intense peaks in each spectrum. The searches of base peak normalized spectra and total ion current normalized spectra are conducted with an absolute value distance metric, while the unit length spectral vectors are compared with a Euclidean distance metric for computational convenience. As with the distance metrics, the choice of a normalization method has little effect on the observed search performance.

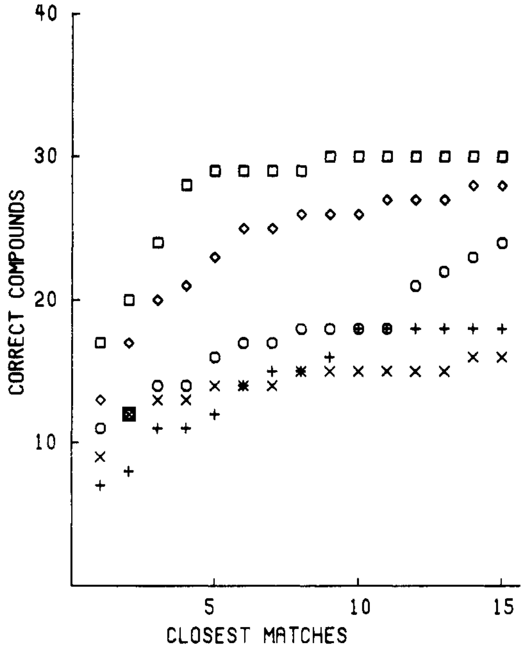


Figure 4. Results for searches with different binary data-encoding methods: ○, complete spectra; +, selected mass positions; ×, 96 bit compressed from complete spectra; ◇, one peak per 14 amu; □, two peaks per 14 amu.

An alternative data-encoding method reduces the amount of intensity information associated with each peak and relies on the mass position information. The use of a peak/no peak (or binary) encoding of mass spectra for library searches was first reported by Grotch.[7] The effectiveness of employing various peak selection methods for encoding binary mass spectra has also been examined.[8,11] Information theory has been used to aid in the compression of binary mass spectra by combining different mass positions[14] and to guide the selection of a limited number of specific mass positions for encoding binary spectra.[15] Results indicating search performance for several methods of encoding binary spectra are reported in Figure 4. In all cases, reference and unknown spectra are compared by counting the number of mismatches between them. "Selected mass position" spectra use only data for the 120 mass positions reported as most useful by van Marlen and Dijkstra. "Compressed binary" spectra consist of 96 bit binary spectra generated according to the method described by Wangen et al. In these spectra some bits correspond to single mass positions while others combine data for several mass positions. For example, a 1 in the 13th bit of such a spectrum indicates the presence of a peak at 45 or 46 amu in the original mass spectrum. Encoding only the single most intense peak every 14 amu is the peak selection method for binary spectra recommended by Grotch. In contrast to the performance observed for spectra with detailed intensity information, the results for complete binary spectra are not the best. This presumably is because a binary encoding forces strong and weak peaks in a spectrum to assume equal importance. The two methods utilizing information theory show results roughly comparable to those results observed for complete spectra while substantially reducing the amount of data stored for each spectrum. However, better results are observed for the methods which select peaks in intervals across a spectrum. Interestingly, if the 96 bit information theory compressed spectra are generated using a library previously abbreviated to contain only the two most intense peaks per 14-amu interval, the search results improve markedly.

Table I. Summary of Distance Metrics for Binary Spectra[a]

Hamming	$D_H = U \text{ XOR } L$
Grotch	$D_G = U \text{ XOR } L - 2\,(U \text{ AND } L)$
Tanimoto	$D_T = \dfrac{U \text{ AND } L}{U \text{ IOR } L}$

[a] XOR = exclusive or; IOR = inclusive or; AND = and.

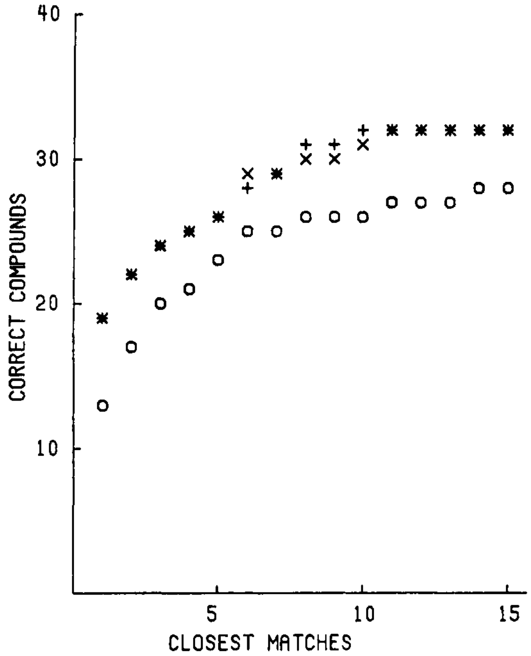


Figure 5. Results for searches with different binary distance metrics: ○, Hamming; +, Grotch; ×, Tanimoto.

Different distance metrics can also be used to compare binary spectra. Grotch has recommended a distance metric that incorporates the number of peaks common to reference and unknown spectra as well as the number of mismatches.[10] This distance metric is the number of mismatches less twice the number of peaks that the spectra have in common. Gray has suggested that other binary metrics may be more discriminating than Grotch's metric.[25] One metric not mentioned by Gray but closely related to those he cites has been used previously in work with binary infrared spectra.[26] These distance metrics are summarized in Table I in terms of Boolean operations. Figure 5 reports the results for searches using these three metrics with binary spectra encoded to retain the single most intense peak in 14-amu intervals. The two distance metrics which employ two Boolean functions show similar results, and both are better than the distance metric based on a single Boolean function. Thus, search performance is improved at the expense of increased computational complexity. Also, the distinct difference in search performance with different peak selection methods was reduced when the more complex distance metrics were used. This effect is illustrated in Figure 6, which shows the results obtained for searches using the Grotch metric with binary spectra encoded by several different methods.

The final type of search algorithm considered here is that which employs an ion series data compression.[16] Such a data compression method offers a unique but practical alternative to peak-oriented data-encoding techniques. Spectra are reduced to a set of n numbers, which indicate the fraction of the total ion current that is attributable to ions occurring in n distinct ion series within a spectrum. Ions appearing at mass positions separated by n-amu intervals belong to the same series and have their intensities added together to produce n sums. These sums are divided by the total of all peak intensities to produce the modulo n ion series compressed spectrum. Figure 7 gives the results obtained with searches using ion series spectra based on 7-, 10-, and 14-amu intervals. Reference and target spectra are compared with an absolute value distance

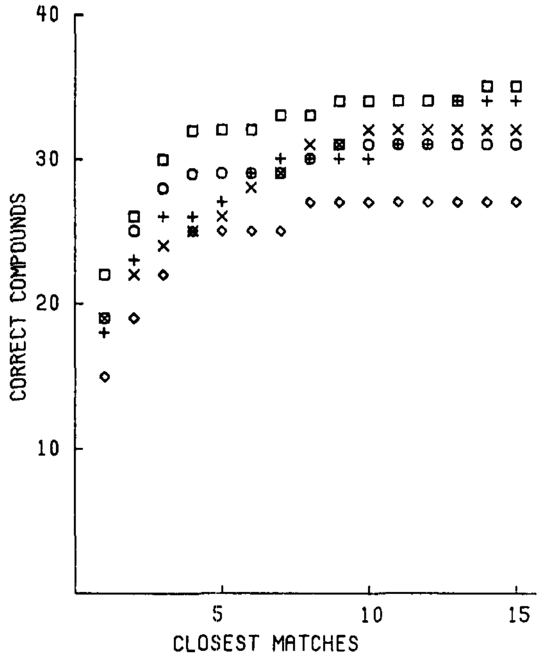


Figure 6. Results for searches with Grotch metric and different data-encoding methods: ○, complete spectra; ◇, selected mass positions; +, 96 bit compressed from abbreviated spectra; ×, one peak per 14 amu; □, two peaks per 14 amu.

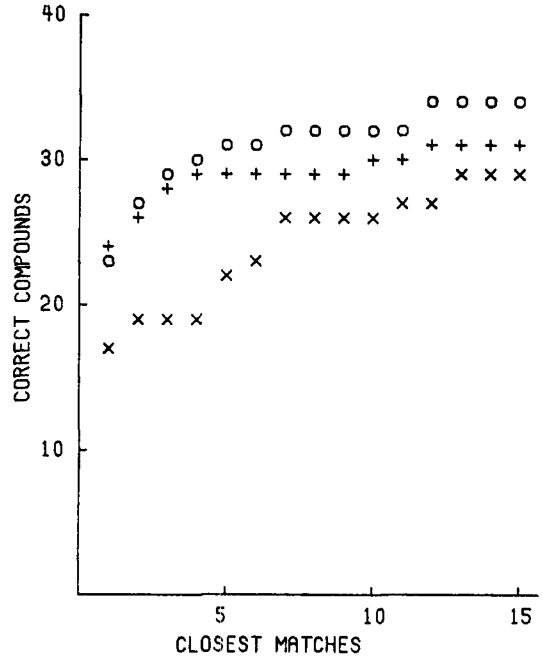


Figure 7. Results for different modulo N spectra: ○, modulo 14; +, modulo 10; ×, modulo 7.

metric. The performance observed for the modulo 14 spectra is comparable to the best results obtained with searches of binary spectra.

Comparisons based on trial searches indicate that mass spectra are best identified by using all the available mass position and intensity information. Reducing the amount of information considered by a distance metric degrades search performance, but some methods of data compression are clearly superior to others. Considerations such as the computer storage requirements for library spectra or search execution time may favor the selection of one data compression method over another. For example, with this library file, a spectrum stored with both mass position and intensity information for a single peak packed into a single 16-bit computer word requires about 50 words on the average. Retaining such information for only the single most intense peaks in 14-amu intervals requires about 12 computer words for the average spectrum. Modulo 14 ion series spectra with an intensity resolution of 0.4% require seven words per spectrum, and binary spectra encoded with one peak per 14-amu interval use about four words for the average spectrum. Thus, storage requirements for spectra encoded in different ways vary by an order of magnitude. A similar variation in search execution times is observed, depending on the data-encoding method and the specific distance metric employed. For searches of binary spectra, use of the Grotch or Tanimoto distance metric probably improves search performance enough to merit the increased computational complexity over the simple Hamming distance metric. However, for searches using detailed intensity information, the choice of a distance metric or normalization makes little difference in terms of search performance and therefore might be guided by considerations such as search speed.

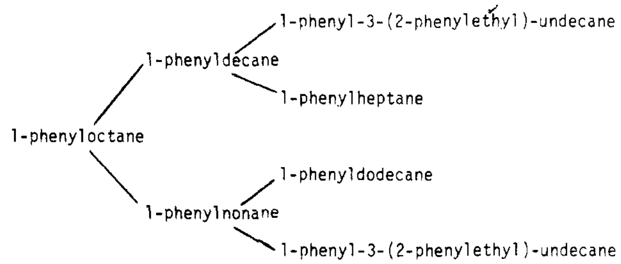


Figure 8. Propagation tree for 1-phenyloctane.

SEARCH EVALUATION BY A PROPAGATION METHOD

In addition to the ability of a search to correctly identify compounds which are represented by library entries, the ability of a search to retrieve similar compounds when the library does not contain a spectrum corresponding to the unknown compound is an important aspect of search performance. The effectiveness of a search system in this respect will depend not only on the search algorithm but also on the library file. Whether a search locates compounds similar to an unknown will inevitably depend on whether similar compounds are members of the library file. The use of a single library file allows the relative performance of different search algorithms to be evaluated. If the file is large and general, the possibility of search results providing a useful interpretative guide for the identification of the unknown is greater than if the file is small and specific. However, the search must then discriminate against a great number of unrelated compounds. The propagation method described above was used to test this feature of search performance.

The results of the first two steps of a propagation are presented in Figure 8. This propagation considers the three nearest matches to a target so that as many as two new compounds will be included with each search. The search algorithm used is based on a Euclidean distance comparison of base peak normalized spectra with information for all peaks included. The seed compound is 1-phenyloctance. With the first step, two new compounds are included in the propagation subset. Upward branches of the diagram indicate the second closest match. The second closest match is always the seed compound because the spectrum is taken directly from the library file. Up to four new compounds could be added with the second step, but only three are in this case. Searches of these three compounds could add six new members to the subset in the next step. Although not shown in Figure 8, only one compound is added with the third step, and by the sixth step no new compounds are introduced so that the propagation

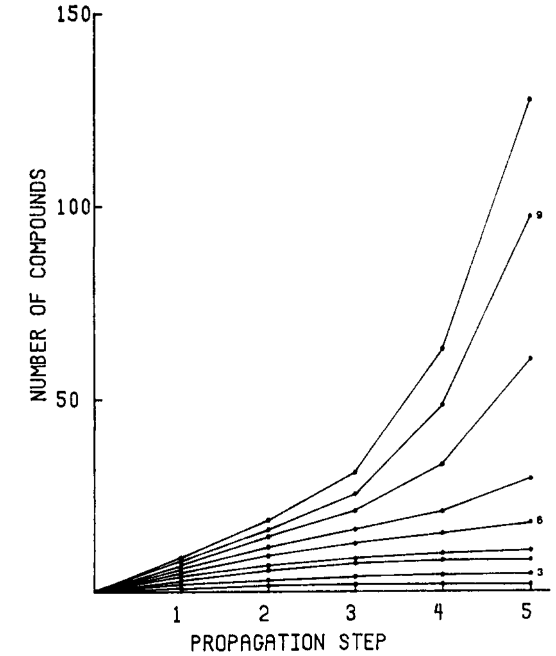


Figure 9. Statistics for propagations of different numbers of nearest matches.

Table II. Compounds Used for Propagation Statistics

1-methylcyclopentene	1,2-dimethylnaphthalene
hexanoic acid	1,1-diethoxypentane
3-thiaheptane	2,4'-dimethylbiphenyl
1-methyl-4-isopropylbenzene	pentobarbital
2,7-dimethyloctane	1-methylchrysene

terminates or closes. This example illustrates the basic features of the propagation method.

An important parameter in a propagation is the number of nearest matches included for repeated searches. The rate at which new compounds are included in a propagation subset and the tendency of subsets to close will depend on the number of nearest matches propagated. Figure 9 is a graph of the number of compounds in a subset vs. the number of propagation steps for propagations which include varying numbers of nearest matches. The search algorithm is again a Euclidean distance search of complete, base peak normalized spectra. Plotted points are mean values for propagations of the ten compounds listed in Table II. Interpolated lines are intended only to indicate which data points are related. When only a few nearest matches are considered, subsets grow very slowly and in many cases close. Even with four nearest matches propagated, nine of the ten subsets contributing to these statistics are closed by the fifth step. However, as more nearest matches are included, more diverse compounds will be forced into the propagation subset causing it to grow rapidly. Subsequent discussion will consider propagations of five or fewer nearest matches.

The rate of growth of a propagation subset will also depend on the search algorithm used, and the relative rates for different algorithms may provide an indicator of relative search performance. Figure 10 is a graph of the number of compounds in a subset vs. the number of steps for propagations of five nearest matches with three different search algorithms. The data points again are means for propagations of the ten compounds listed in Table II. Besides the previously described Euclidean distance search, a search of modulo 14 spectra compared with an absolute value distance metric and a search of compressed binary spectra generated from Biemann abbreviated mass spectra and compared with the Tanimoto distance metric were used to produce propagation subsets. Both of these latter search algorithms propagate at a significantly higher rate than did the Euclidean distance search. The average rate for the search of modulo 14 spectra is comparable to that for a propagation with the Euclidean distance search which includes seven nearest matches. Similarly, the rate for the binary search is comparable to the rate for a Euclidean search propagation which includes the nine nearest matches. If a lower propagation rate is taken as an indication of better search performance, these propagation statistics rank these three search algorithms in the same order as do the results for trial searches of target spectra.

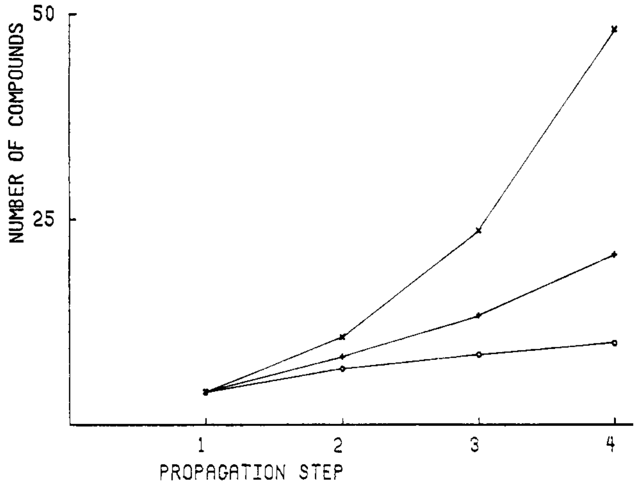
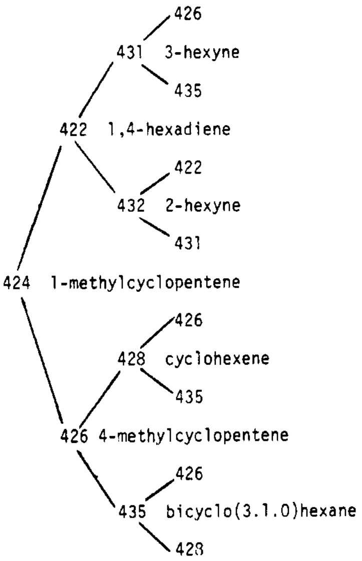


Figure 10. Statistics for propagations of five nearest matches with different search algorithms: O complete spectra; +, modulo 14; ×, binary.



Compounds are listed with serial numbers and names at their first appearance.

Figure 11. Propagation tree for a search of compressed binary spectra with 1-methylcyclopentene as the seed compound.

However, instead of general statistics, results for specific propagation tests may provide sufficient information for the evaluation of different search algorithms. Figure 11 shows the propagation tree for three nearest matches as found by the search of compressed binary spectra beginning with 1-methylcyclopentene as the seed compound. No new compounds are introduced in the third step, and six unique compounds are included in the subset. All six compounds have the same molecular formula, and none is particularly out of place in the subset. Figure 12 reports the results for a propagation of the same seed compound with a search of modulo 14 spectra. In this case, the propagation closes after

Table III. Propagation Subsets for Three Searches with 1,1-Diethoxypentane as the Seed Compound and Five Nearest Matches Considered

	complete spectra	modulo 14 spectra	96 bit compressed spectra
first level	1,1-diethoxyisopentane 1,1-diethoxy-2-methylbutane 1,1-diethoxyoctane 1,1-diethoxynonane	1,1-dimethoxybutane 1,1-diethoxybutane 1,1-diethoxyisopentane 1,1-diethoxyhexane	1-pentanal imidazolidin-2-one 1,1-diethoxypropane 1,1-diethoxyisopentane
second level	triethoxymethane 1,1-diethoxy-3-methyl-3-butene 1,1-diethoxy-2-hexene 1,1-diethoxyheptane 1,1-diethoxydecane 1,1-diethoxy-10-undecene	1,1-diethoxy-2-methylpropane	5-aminotetrazole 2-methylbutanal 3-methylbutanal ethyl 2-propenyl ether butyrolactone methyl-*n*-butylamine 1-methoxy-2-propanol *N*-methylpiperidine *n*-valeramide 1,1-dimethoxypropane 1,1-diethoxy-2-propene bis(1-methyl-2-hydroxypropyl) ether 1-(1,3-dimethylbutoxy)-2-propanal 1,1-diethoxy-2-methylbutane
third level	1,1-diethoxyhexane 2 other compounds	1,1-diethoxy-2-methylbutane	1,1-diethoxybutane 36 other compounds

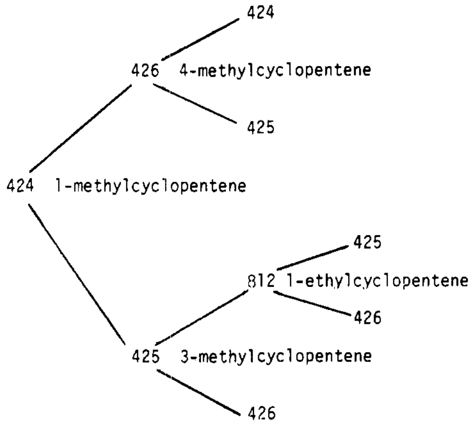


Figure 12. Propagation tree for a search of modulo 14 spectra with 1-methylcyclopentene as the seed compound.

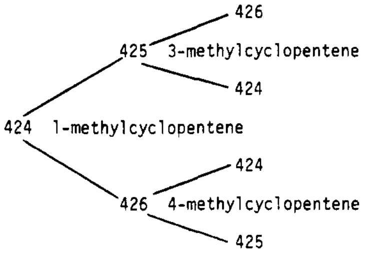


Figure 13. Propagation tree for a search of complete spectra with 1-methylcyclopentene as the seed compound.

two steps with the addition of only three new compounds to the subset. Two compounds are the other geometrical isomers of the seed compound, and the third is a homologous compound. While the propagation subset for this search has included a compound with a molecular formula different from that of the seed compound, it did not include any dienes, alkynes, or bicycloalkanes, nor did it omit one of the geometrical isomers of the seed compound. To complete the comparison, Figure 13 shows the propagation tree obtained with the Euclidean distance search. The subset closes after the first step with the inclusion of the three methylcyclopentene isomers. This search algorithm finds the three isomeric compounds in a tighter cluster than do the other two searches. Taken together these three propagation trees, which represent a total of 14 individual search runs, give a reasonable impression of the relative performance of the three search algorithms.

As more than three nearest matches are considered, it becomes increasingly difficult to draw propagation trees, but the relevant information can be effectively summarized by listing the new compounds introduced into a propagation subset at each step of a propagation. Table III lists the results for a propagation of five nearest matches using the same three search algorithms and beginning with 1,1-diethoxypentane as the seed compound. The search of binary spectra locates two other 1,1-diethoxyalkanes among the four new nearest matches in the first step. The second step adds 14 new compounds including two more 1,1-diethoxyhydrocarbons, and the third step adds 37 new compounds which include one more 1,1-diethoxyalkane. With this search algorithm, the propagation subset has diverged rapidly. By contrast the subsets for propagations with the other two search algorithms are more homogeneous. The search for the modulo 14 spectra includes one 1,1-dimethoxy- and three 1,1-diethoxyalkanes in the first step, and one more 1,1-diethoxyalkane in each of the next two steps. The Euclidean search of complete spectra finds four 1,1-diethoxyalkanes in the first step, including two isomers of the seed compound. The second step of this propagation introduces five more 1,1-diethoxyhydrocarbons and triethoxymethane. After the third step the propagation subset includes a total of 13 new compounds of which ten are 1,1-diethoxyalkanes or -alkenes. If the propagation for the modulo 14 search is continued for five steps, the subset of new compounds grows to 15, 9 of which are 1,1-diethoxyalkanes. The library file contains a total of 15 1,1-diethoxyhydrocarbons counting the seed compound. Thus, for this seed compound, about 60% of the homologous compounds in the library file are located in the first few steps of the propagation performed with the search of modulo 14 spectra, and about 80% of them are found with the search of complete spectra.

The propagation method can sometimes show differences between search algorithms which give very similar results when compared with trial searches. For example, if three rather than five nearest matches are propagated for 1,1-diethoxypentane with the Euclidean distance search, no new compounds are introduced after the first step, and the closed subset consists of the three isomeric 1,1-diethoxyalkanes with the molecular formula $C_9H_{20}O_2$. Using the Biemann search algorithm for the same propagation produces a subset which is closed after five steps and has included 11 1,1-diethoxyhydrocarbons.

These results are typical of the performance observed for the representative search algorithms tested with the propagation method. Similar results are obtained with a wide variety of seed compounds, including aldehydes, substituted aromatic compounds, and barbituric acid derivatives. The propagation results give an indication of the relative ability of different mass spectral search algorithms to select compounds that are chemically similar to an unknown. While propagation rate statistics can be taken as an indicator of relative search performance, a simple examination of the compounds contained in a propagation subset after only a few steps can be quite informative.

CONCLUSIONS

The relative performance of mass spectral search algorithms has been evaluated by two methods. Trial searches with target compounds are useful in assessing how well a search algorithm can identify mass spectra through the retrieval of matching library entries. The propagation of selected spectra, by repeated searches of nearest matches, provides an indication of how effectively search algorithms can locate spectra of related compounds. Testing a variety of search algorithms by these complementary methods indicates that the searches which perform best use complete mass position and intensify information. Some data-encoding methods significantly impair search performance. Other methods can achieve substantial reductions in the amount of information stored and processed by the search algorithm while maintaining a respectable level of performance. The final selection of a search algorithm is probably best made by the individual analyst, who is aware of the relative advantages and limitations of the different methods available and can relate these to the specific problem at hand. The propagation technique offers a new and useful approach to the evaluation of search performance.

ACKNOWLEDGMENT

The authors would like to acknowledge the contributions through helpful discussions of J. C. Marshall and S. R. Lowry. This research has been supported by the National Science Foundation.

REFERENCES AND NOTES

(1) P. D. Zemany, *Anal. Chem.*, **22**, 920 (1950).
(2) F. W. McLafferty and R. S. Gohlke, *Anal. Chem.*, **31**, 1160 (1959).
(3) S. Abrahamsson, S. Stallberg-Stenhagen, and E. Stenhagen, *Biochem. J.*, **92**, 2p (1964).
(4) S. Abrahamsson, *Sci. Tools*, **14**, 29 (1967).
(5) B. Pettersson and R. Ryhage, *Ark. Kemi*, **26**, 293 (1967).
(6) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1465 (1968).
(7) S. L. Grotch, *Anal. Chem.*, **42**, 1214 (1970).
(8) B. A. Knock, J. C. Smith, D. E. Wright, and R. G. Ridley, and W. Kelly, *Anal. Chem.*, **42**, 1516 (1970).
(9) H. S. Hertz, R. A. Hites, and K. Biemann, *Anal. Chem.*, **43**, 681 (1971).
(10) S. L. Grotch, *Anal. Chem.*, **43**, 1362 (1971).
(11) S. L. Grotch, *Anal. Chem.*, **45**, 2 (1973).
(12) R. J. Mathews and J. D. Morrison, *Aust. J. Chem.*, **27**, 2167 (1974).
(13) R. J. Mathews, *Int. J. Mass Spectrom. Ion Phys.*, **17**, 217 (1975).
(14) L. E. Wangen, W. S. Woodward, and T. L. Isenhour, *Anal. Chem.*, **43**, 1605 (1971).
(15) G. van Marlen and A. Dijkstra, *Anal. Chem.*, **48**, 595 (1976).
(16) G. T. Rasmussen, T. L. Isenhour, and J. C. Marshall, *J. Chem. Inf. Comput. Sci.* **19**, 98 (1979).
(17) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972).
(18) F. W. McLafferty, R. H. Hertel, and R. D. Villwock, *Org. Mass Spectrom.*, **9**, 690 (1974).
(19) F. P. Abramson, *Anal. Chem.*, **47**, 45 (1975).
(20) J. A. de Haseth, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, *Anal. Chim. Acta Comput. Tech. Optm.*, **103**, 109 (1978).
(21) K.-S. Kwok, R. Venkataraghavan, and F. W. McLafferty, *J. Am. Chem. Soc.*, **95**, 4185 (1973).
(22) E. Stenhagen, S. Abrahamsson, and F. W. McLafferty, "Registry of Mass Spectral Data", Wiley-Interscience, New York, 1974.
(23) R. M. Silverstein and G. C. Bassler, "Spectrometric Identification of Organic Compounds", 2nd ed, Wiley, New York, 1967.
(24) H. W. Brown and E. J. Bonelli, *Abstr., 1977 Pittsburgh Conf. Anal. Chem. Appl. Spectros.*, 144 (1977).
(25) N. A. B. Gray, *Anal. Chem.*, **48**, 1420 (1976).
(26) H. B. Woodruff, S. R. Lowry, G. L. Ritter, and T. L. Isenhour, *Anal. Chem.*, **47**, 2027 (1975).