Registry Number: 102-54-5
Molecular Formula: $C_{10}H_{10}Fe$

Figure 8. Ferrocene.

of database integrity, providing for symmetrical substance registration, and maintaining direct access file addresses.

Within the CAS Chemical Registry System, the ACMF is used to support database integrity by retrieving for chemists' review those substances new to the CAS Chemical Registry database which have an ACMF equal to substances already registered in the CAS database. This approach is valid since the ACMF calculation is independent of the unique table generation algorithm. The use of a manual review is practical since there are so few ambiguous ACMF's, i.e., distinct substances with identical ACMF's.

The computer time requirements for the unique table generation algorithm to uniquely label this small number of highly symmetrical substances is prohibitively large. Ferrocene, shown in Figure 8, is an example of a structure which would require 10! or 3,628,800 iterations in order to be uniquely labeled. For substances of this type, which cannot be uniquely labeled within a reasonable amount of computer time, the ACMF is utilized to determine if the substance is new to the CAS Chemical Registry database. The ACMF is utilized to identify file substances which are potentially identical with the candidate substance. For these highly symmetrical substances, final determination of whether the candidate substance is new is made via computer-based atom-by-atom structure comparison of the candidate substance with those file substances identified as potentially identical by the ACMF.

Finally, since the CAS Chemical Registry System utilizes a direct access database, it is necessary to have a precise file address based on structural properties. This is provided by the ACMF.

As noted earlier, the ACMF is not always successful in distinguishing chemical substances sharing the same molecular formula. Since the initial installation of the CAS Chemical Registry System, minor improvements have been made to the ACMF algorithm, but the problems illustrated with the substances in Figure 7 are more deeply rooted and a solution to these problems will require a substantially more complex algorithm. However, since the ACMF fails in so few cases, it provides a practical tool for use with files of chemical substances. In addition, it is a failsafe mechanism because substances are added to the CAS Registry database only after applying the unique table generation algorithm.

ACKNOWLEDGMENT

The development of the CAS Chemical Registry System was substantially supported by the National Science Foundation (Contract C656). Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

REFERENCES AND NOTES

(1) O'Korn, L. J. "Algorithms in the Computer Handling of Chemical Information", ACS Symposium Series No. 46, "Algorithms for Chemical Computations"; American Chemical Society: Washington, D.C., 1977; p 122.
(2) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-124.
(3) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107-113.
(4) Connectivity values are numbers associated with each atom of a substance which are derived as part of the unique table generation algorithm. They have been used in the CAS Chemical Registry System since 1965. These connectivity values reflect the atom-by-atom interconnections but ignore the element value and bond type. The adjective "augmented" indicates the inclusion of the element value and bond-type information in the computation of the connectivity values.
(5) Dittmar, Stobaugh, and Watson have provided descriptions for these special structural characteristics.

Mass Spectral Library Searches Using Ion Series Data Compression

GREGORY T. RASMUSSEN and T. L. ISENHOUR*

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

JOHN C. MARSHALL

Department of Chemistry, Saint Olaf College, Northfield, Minnesota 55057

Received October 27, 1978

A series of library searches using a mass spectral data compression method based on fractional ion currents of specific ion series is described. The method offers efficient data compression, reasonable search performance, and a capability for use with file partitioning to reduce search times.

Since its initial development, the computerized search of mass spectral data has gained widespread acceptance as a useful tool for the identification of diverse compounds. The variety of mass spectral search systems reported in the literature and the automatic inclusion of at least one library search program in computerized data systems now commercially available for mass spectrometers attest to the importance of the technique. Computerized searches require that mass spectral data for every compound in a reference library be retained on some storage device peripheral to the computer, such as a magnetic tape or a disk pack. The growth of reference collections of mass spectral data to the extent that such collections routinely contain data for tens of thousands of compounds implies that a large amount of computer-ac-

0095-2338/79/1619-0098$01.00/0

cessible peripheral storage space is necessary in order to use this data in a library search system. Alternatively, efficient methods for the reduction of the amount of mass spectral data retained for each compound can be applied to limit the amount of storage space required for the library data file. These methods of reducing the space requirements of a library file are referred to as data compression techniques. A second important benefit of data compression is that computer searches of compressed library files can be run more rapidly than searches using uncompressed library data. Because less data is associated with each mass spectrum in a compressed file, less data need be considered by the search's comparison algorithm, and search speed is increased. The goal of efficient data compression is to eliminate information that is not useful in characterizing the mass spectrum of a compound while retaining enough information to distinguish that mass spectrum from those of other compounds.

PEAK-ORIENTED MASS SPECTRAL DATA COMPRESSION

Data compression efforts generally begin by selectively ignoring some mass spectral information. Digitized mass spectra consist minimally of a pair of numbers for each peak observed. These two numbers indicate the mass-to-charge ratio or mass position and the relative intensity of each peak. Essentially all mass spectral search algorithms reported to date rely on specific information about the mass positions of peaks in the library and unknown spectra. One approach to data compression is to process mass spectra so that only certain peaks are retained. This is done by selecting the most intense or otherwise most "significant" peaks from either the entire mass spectrum or from segments of the spectrum. Some searches based on selected peaks use algorithms that require explicit intensity information. For example, an early ASTM data collection listed the mass positions and intensities of the six most intense peaks per spectrum, and this compression scheme has been the basis of computerized searches.[1-4] A compression method which retains the mass position and intensity of the two most intense peaks in 14-dalton (amu) segments of the mass spectrum was employed in the search algorithm developed by Biemann and his co-workers.[5] Also, criteria other than peak intensity have been used to determine which peaks are retained, as in the recently reported search algorithm in which the ten most "significant" peaks in each spectrum are kept, with "significance" being defined as the product of a peak's intensity and its mass position.[6] Another major approach to data compression is to retain all peaks with intensities above a threshold but to limit the intensity information stored for each peak. The extreme case with this approach is represented by the binary mass spectrum in which the intensity resolution is reduced to the extent that only a peak's presence or absence is considered by the search algorithm.[7,8] The data compression obtainable with binary mass spectra arises because only a single bit of a computer word is required to store the information associated with each mass position in a spectrum. Search algorithms which combine these approaches by comparing only the mass positions of selected peaks have also been devised. For example, Knock et al. have studied data compressions which retain the n most intense peaks in intervals of m daltons, where n and m are integers.[9] One popular compression of this type, which uses the mass position of the single most intense peak in 14-dalton intervals, allows data for each interval of the spectrum to be stored in four bits of a computer word.[8] An approach to data compression in which detailed information about the mass positions of peaks in the original spectrum is sacrificed is illustrated by the "reduced dimensionality" binary spectra of Isenhour and his co-workers. With this method a single bit of a computer word indicates the presence of a peak at one or more of several different mass positions. This technique attempts to represent binary mass spectra with peaks at masses up to several hundred daltons using substantially fewer bits, with the combinations of mass positions being selected according to the principles of information theory.[10]

ION SERIES DATA COMPRESSION

All of these data compression techniques are distinctly peak oriented because they all require more or less specific information about the mass positions of peaks in the original spectra. A contrasting approach to the compression of mass spectral data is to represent spectra as fractional "ion currents" associated with 14 ion series within a spectrum. Such representations of mass spectra, which were first described by Hamming and Grigsby, have been employed for purposes of compound classification in two contexts.[11] One pattern of this type, referred to as a "rectangular array", plays a role as a prefilter in the search algorithm developed by Biemann and co-workers. The prefilter is intended to select only those library spectra which are reasonably similar to an unknown for detailed comparison by the search algorithm and is described as a "rough screen for compound type".[5] Crawford and Morrison have illustrated the potential for use as compound classifiers of similar 14-dimensional patterns, which they called "reduced mass spectra".[12] Independently, Smith, who referred to his patterns as "ion series summation spectra", developed 50 composite patterns, each characteristic of a specific chemical functional group or molecular feature, and used these to predict the molecular class of unknown compounds.[13]

These data treatment methods are similar in that none retains mass position or intensity information about specific peaks in the original spectra. In each case, the mass position and intensity information is used to produce a set of 14 numbers which represent the original mass spectrum. For convenience, this set of numbers can be thought of as a 14-dimensional vector or pattern. The basic method of Smith and of Crawford and Morrison was to identify a molecular class, to select the mass spectra of all compounds belonging to that class from a library collection, to compute a 14-dimensional pattern from each of these spectra, and then to compute a composite pattern by averaging the individual patterns of all class members. Once a set of composite patterns has been generated, classification of an unknown compound is performed by computing a 14-dimensional pattern from the unknown mass spectrum and then calculating a distance or dissimilarity between the pattern of the unknown and each composite pattern. The unknown compound is classified as a member of the molecular class represented by the composite pattern that is most similar to the unknown pattern. Instead of using an ion series pattern for compound classification, the pattern of an individual compound can be used directly as a compressed form of the mass spectrum. A library search based on such a compression would simply require computation of an ion series pattern from the mass spectrum of an unknown compound and then comparison of this pattern with a reference library of previously generated ion series patterns.

MASS POSITION MODULO N DATA COMPRESSION

Although the name "ion series summation spectra" is more descriptive than "reduced mass spectra", it seems important to indicate the interval of separation between mass peaks in a related series, which is also the number ion series considered. Therefore, we refer to these patterns as "mass position MODULO 14 summation spectra", or briefly "MOD14 spectra". The representation of full mass spectra as MOD14 spectra assumes the existence of 14 distinct ion series within a mass spectrum. Each peak in a mass spectrum is associated

with a specific ion series depending on its mass position. Peaks separated by 14 daltons (e.g., 44, 58, 72, . . .) are members of the same ion series. The significance of the number *14* is that it corresponds to the total atomic mass of a methylene group. Mass spectra frequently exhibit series of peaks at 14-dalton intervals, and these peaks typically correspond to ions which differ in composition by single methylene groups. A MOD14 spectrum is generated by summing the intensities of all peaks in a given ion series. This is repeated for each ion series to produce 14 sums. Each of these is divided by the sum of the intensities of all peaks in the mass spectrum. The sum of all peak intensities reflects the total ion current, and division of each ion series sum by the total spectrum intensity normalizes the ion series sums so that each indicates the fraction of the total ion current attributable to ions of that series.

All previous investigations of ion series spectra have been based on intervals of 14 daltons and the corresponding 14-dimensional patterns. In evaluating the utility of ion series spectra as an approach to mass spectral data compression, it is interesting to consider the effects of selecting other ion series intervals. In addition to MOD14 spectra, MOD10 and MOD7 spectra have been included in this study. The use of 7- and 10-dalton intervals to arbitrarily define new ion series within a mass spectrum allows representation of mass spectra with only 7 and 10 numbers, respectively. The interval of 7, as a submultiple of 14, produces a representation of the mass spectrum that corresponds to halves of the MOD14 spectrum summed together pairwise. The interval of 10 represents a compression roughly midway between those using intervals of 7 and 14. Comparing results with a MOD10 compression is interesting because it is an arbitrary compression of mass spectra without the benefit of a theoretical basis in terms of the known fragmentation patterns of organic molecules. Taken together, the use of search algorithms based on 7-, 10-, and 14-dimensional patterns should provide a good indication of the effectiveness of the "mass position MODULO N summation" as an approach to mass spectral data compression.

If the different ion series within a spectrum are numbered 1 through N, a simple algorithm for the computation of MODN spectra can be defined. Each peak in a mass spectrum is considered in sequence and the appropriate ion series for the peak is identified by a pointer p, which is computed by adding 1 to the remainder of the division of the peak's mass position by N. (The "Remaindering" operation is identical with the MODULO function used in computer programming; e.g., for division by 14 the remainders are integers 0 to 13.) Once a peak is identified with the proper ion series, its intensity is added to the associated ion series sum. The total ion intensity is simply the sum of all peak intensities or the sum of all ion series intensities. The algorithm is summarized in eq 1–4, where p is the pointer, N is the number of ion series, n is the number of peaks in a spectrum, n_p is the number of those peaks associated with the pth ion series, M_j and I_j are the mass position and the intensity, respectively, of the jth peak, S_p' is the raw sum of ion intensities for the pth ion series, T is the total ion intensity, and S_k is a normalized ion series sum. For

$$p = 1 + \text{MODULO}(M_j,N) \tag{1}$$

$$S_p' = \sum_{j=1}^{n_p} I_j(p) \tag{2}$$

$$T = \sum_{j=1}^{n} I_j = \sum_{k=1}^{N} S_k' \tag{3}$$

$$S_k = S_k'/T \tag{4}$$

MOD14 spectra, this algorithm produces patterns identical with those of Crawford and Morrison. Smith used a slightly different algorithm which resulted in a different numbering of the 14 series sums. If only peaks at mass positions above 30 daltons are considered and if the pointer p^* is calculated according to eq 5, patterns identical with those of Smith can

$$p^* = 1 + \text{MODULO}(M_j - 30,N) \tag{5}$$

be computed. It may be noted that division by the sum of all peak intensities, while necessary for standard base peak normalized spectra, is not required if the original spectra are normalized by the total ion current.

DISTANCE METRICS

Given a set of these MODN spectra for all compounds in a reference library, it is necessary to identify a distance metric that can be used to compare patterns. If the MODN spectra are treated as multidimensional patterns with each ion series representing one of N different Cartesian axes, a single spectrum can be thought of as a point in an N-dimensional space. A common distance metric for the comparison of points in multidimensional spaces is the Euclidean distance, which represents the length of the straight line joining two points in a Euclidean space. The Euclidean distance, which is the metric customarily used in "nearest-neighbor" classification or clustering algorithms, is given in eq 6, where U identifies the

$$D_E = (\sum_{i=1}^{N} [S_{Li} - S_{Ui}]^2)^{1/2} \tag{6}$$

MODN spectrum of the unknown and L signifies a library entry.

A second distance metric which has been used to compare mass spectra and which was used by Smith in his work with compound classifiers is given in eq 7. This distance metric

$$D_A = \sum_{i=1}^{N} \text{absolute value } [S_{Li} - S_{Ui}] \tag{7}$$

will be referred to as the absolute value distance. It corresponds to shortest total length of a series of line segments joining two points in a space if only segments parallel to the reference axes are allowed. The two distance metrics are similar with the absolute value distance giving equal weight to all differences, while the Euclidean distance gives greater weight to greater differences.

EXPERIMENTAL SUMMARY

The data set used for this work consisted of the mass spectra of 16 924 organic compounds drawn from a larger set of mass spectra collected by McLafferty and his colleagues. Spectra of compounds containing only the elements carbon, hydrogen (including deuterium), oxygen, nitrogen, sulfur, phosphorus, and the halogens were collected in a data file. All peaks occurring at integral mass positions and having intensities above 1% were included, producing a data set with a total of more than 816 000 peaks at masses up to nearly 660 daltons. MOD14, MOD10, and MOD7 spectra for each compound were computed as described above and kept in separate random access library files. The three sets of MODN spectra were also computed from 20 mass spectra selected from a compilation of data for biochemically significant compounds and from the 51 mass spectra given with problems in the last three chapters of the text by Silverstein and Bassler.[14,15] These MODN spectra were treated as target or "unknown" spectra and searched with their respective library files of MODN spectra using both Euclidean and absolute value distance metrics. Additionally, computer searches based on the complete original mass spectra and on binary mass spectra were run for comparison. Also, the 50 composite patterns published by Smith were used as target patterns for searches

with the MOD14 and MOD7 libraries. All computer programs were written in Fortran IV and run on an IBM 360/75 computer operating at the University of North Carolina Computation Center.

RESULTS AND DISCUSSION

Three factors must be considered when assessing the effectiveness of a method for mass spectral data compression. The efficiency of the data compression is indicated by the amount of computer-accessible storage space necessary to contain all data for a given library as compared with the space required by alternative compression methods. Second, the performance of a search algorithm as evidenced by its ability to reliably identify unknown compounds must be consistent with the needs of the user. This can be evaluated through the use of target spectra drawn from a source other than the library file. Finally, the time required to perform a search should be reasonably low and data compression methods that allow rapid searching are to be preferred.

COMPRESSION EFFICIENCY

In order to evaluate the efficiency of the MODN compressions, computer storage requirements were determined for the original data set, three conventional compression techniques, and the MODN compressions. For a given data set, space required to store information such as compound names, molecular weights, and molecular formulas is constant and independent of the form of the mass spectral information and therefore is not included in these calculations. For complete mass spectra and for the compression methods based on selected peaks, stored data will consist of mass position and intensity values for each peak. If the stored spectra contain variable numbers of peaks, a peak counter is also required. Allowing a 16-bit computer word for each mass position, intensity value, and peak counter results in a total storage requirement of nearly 1.65 megawords for the full data set. If only the two most intense peaks in 14-dalton intervals are stored, as in the Biemann abbreviation, the required storage is slightly over 0.70 megaword. Likewise if only ten peaks are retained for each library entry, the storage requirement is approximately 0.34 megaword. If an intensity resolution of 1% is deemed sufficient and if no compounds contain peaks at mass positions above 654 daltons, the mass position and intensity information for each peak can be packed into a single 16-bit computer word, reducing the storage requirements for these data files by a factor of 2. However, such packing of words requires some computational effort to decode the data when it is read so that computer time is traded for storage space. The binary data compression allows data for 16 mass positions to be represented in one computer word. The conventional binary encoding of mass spectra will be most efficient if variable-length binary spectra are stored with each spectrum having only as many words as are necessary to include the peak at the highest mass position. For this library of mass spectra, such a binary data file requires nearly 0.26 megaword of storage space. A MOD14 spectrum consists of 14 numbers which can be stored as a series of integers scaled to sum to a constant, such as 10000. Allowing 14 words for each MOD14 spectrum yields a total storage requirement of less than 0.24 megaword. Similarly, a MOD10 compression requires 0.17 megaword, and a MOD7 file occupies less than 0.12 megaword of storage space. If the data for the MODN compressions are scaled to 250, which allows an intensity resolution of 0.4%, data for two MODN dimensions can be stored in a single word. This cuts the storage requirements for the MODN spectra in half at the expense of the computer time necessary to unpack the pairs of data. Relative storage requirements for these data compression techniques, in terms of computer words per average spectrum without bit packing, is illustrated in Figure 1. The use of bit packing would halve the requirements for all storage methods except the binary spectra. Using one computer word per dimension, the MOD14 compression is slightly more efficient than the binary encoding and needs only one-seventh the storage space of the complete mass spectral data file.

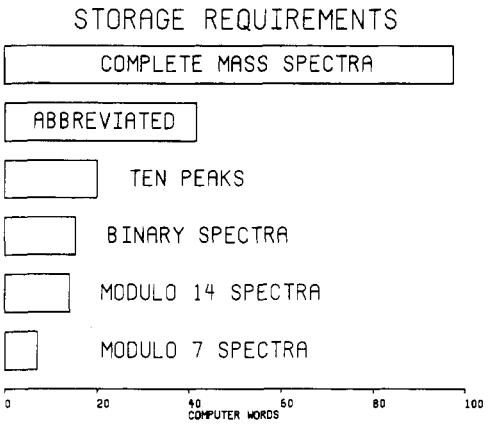

Figure 1. Bar graph showing relative storage requirements per spectrum for various data compression techniques.

SEARCH PERFORMANCE

As a preliminary test of the performance of a library search based on the MOD14 and MOD7 compressions, the 50 composite patterns published by Smith were searched as target patterns using both the Euclidean distance and the absolute value distance metrics. Because each composite pattern represents a molecular class, the compounds selected as nearest matches are expected to belong to the molecular class of the target pattern. Results obtained using the Euclidean distance metric with MODN spectra were generally similar to those obtained with the absolute value distance metric, although the order of nearest matches sometimes differed. The Euclidean distance search would sometimes fail to include compounds of the correct molecular class which were selected as nearest matches by the absolute value search, and on this basis the latter distance metric was judged to be slightly superior. Table I lists the five nearest matches for five different molecular classes as found by MOD14 and MOD7 searches using the absolute value distance metric. For most composite patterns compounds of the correct molecular class predominate among the nearest matches with the MOD14 results being noticeably better than those for the MOD7 search. For example, the 15 nearest matches to the general aliphatic ketone pattern are aliphatic ketones with both MOD14 and MOD7 searches. However, with the pattern for substituted two ring polynuclear aromatics, 12 of the 14 nearest matches selected by the MOD14 search are substituted naphthalenes, whereas only five of the 14 nearest MOD7 matches are of the correct molecular class. The results in Table I show an occasional anomalous compound among the nearest matches for some molecular classes, and for some composite patterns compounds of the correct molecular class are only scattered throughout the nearest matches. This relatively poorer performance is generally observed for those composite patterns based on relatively fewer mass spectra and those identifying rather specific structural features. Overall, the results indicate that the MOD14 compression and, to a lesser extent, the MOD7 compression do preserve molecular structure information contained in mass spectra.

To further investigate the effectiveness of the MODN searches, two sets of mass spectra of individual compounds were converted to MODN spectra and searched as target

Table I. Five Nearest Matches to Each of Five Composite Patterns Selected by the MOD14 and MOD7 Searches

molecular class	MOD14	MOD7
normal alkanes	*n*-octadecane	*n*-octadecane
	n-pentadecane	*n*-tetracosane
	n-heptadecane	*n*-pentadecane
	n-tetracosane	*n*-heptadecane
	n-hexadecane	*n*-tridecane
n-alk-1-yne	1-octyne	1-octyne
	1-decyne	1-decyne
	1-nonyne	1-dodecyne
	1-dodecyne	1-heptyne
	1-heptyne	3,3,5-trimethyl-cyclohexanol
2 ring-substituted polynuclear aromatics	1,2-dimethyl-naphthalene	2-ethyl-naphthalene
	2-ethyl-naphthalene	1,2-dimethyl-naphthalene
	1-ethyl-naphthalene	1-ethyl-naphthalene
	2,3-dimethyl-naphthalene	1,3-dimethyl-naphthalene
	O-phenylanisole	*O*-chloro-*p*-dodecylphenol
n-alkyl cyclohexane	1-methyl-1-cyclohexylpropane	1-methyl-1-cyclohexylpropane
	1-cyclohexyloctane	1-cyclohexylbutane
	1-cyclohexylbutane	1-cyclohexyloctane
	1-cyclohexyl-2-methylpropane	1-cyclohexyl-2-methylpropane
	1-cyclohexyl-propane	1-cyclohexyl-propane
secondary and tertiary alkyl amines	di-*n*-propylamine	di-*n*-propylamine
	triisoamylamine	triisoamylamine
	diisohexylamine	dodecylphenol
	tri-*n*-propylamine	diisohexylamine
	di(2-ethylbutyl)-amine	di(*n*-hexyl)amine

Table II. The Twenty Compounds in Target Set 2

acetylsalicylic acid	griseofulvin
adenine	*p*-hydroxycinnamic acid
p-aminobenzoic acid	indole
amobarbital	methyl 3,4-dimethoxyphenylacetate
androsterone	methyl palmitate
anthracene	meprobamate
caffeine	morphine
p-cresol	nicotinamide
dibutylnitrosamine	riboflavin
dioctyl phthalate	saccharin

patterns. One consisted of the 51 previously described spectra from Silverstein and Bassler's textbook. This set includes compounds which contain a wide variety of molecular functional groups and structural features and which have molecular weights up to nearly 250 daltons. Forty of these compounds have corresponding spectra in the library file. The second set consisted of 20 spectra for compounds known to be represented in the library file. Because some contributors have spectra appearing both in the compilation that was the source of this set and in the compilation used as the library file, special care was taken to ensure that no spectra identical with those in the library were included in the target set. The compounds in the second set are listed in Table II. The second target set includes some more complex compounds and is expected to provide a more stringent test of the search algorithms. In addition to searches with MODN spectra, searches of full mass spectra using a Euclidean distance metric and of binary mass spectra using a metric proposed by Grotch[16] were conducted with each set of targets. The results for the search using complete mass spectra represent the best performance for any search algorithm using the available data. The results obtained with the binary spectra illustrate the performance of a search based on a data compression with

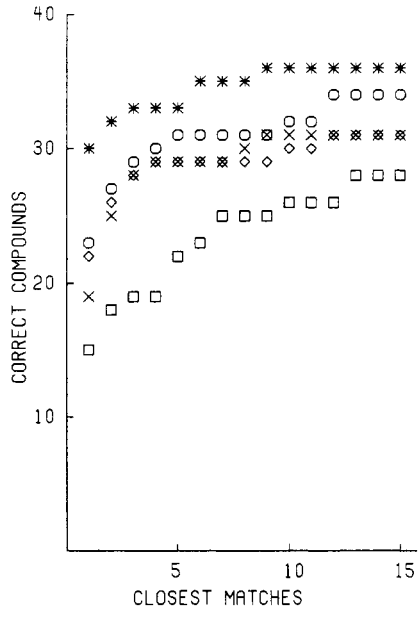

Figure 2. Results of searches on target set 1 for five data compression techniques: (*) uncompressed, (×) binary, (○) MOD14, (◇) MOD10, (□) MOD7.

storage requirements roughly comparable to those of the MOD14 compression.

Search results obtained with the first set of target compounds are summarized in Figure 2, which is 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. These results for the MODN searches reflect the use of the absolute value distance metric, which was slightly superior to the Euclidean distance metric. With the Euclidean distance metric, one or two fewer targets were correctly identified within the 15 nearest matches for a given MODN search. As expected, the search using complete mass spectral data performs best; however, four targets are not correctly identified within the 15 nearest matches. The performance of the MOD14 search is nearest to that of the search based on complete mass spectra, and the MOD10 search results are most nearly comparable to those obtained with the binary search. For the MODN searches, the observed performance declines somewhat as the degree of compression increases.

This method of presenting the search results suffers the limitation that no consideration is given to the appearance of geometric isomers or homologous compounds among the nearest matches. With the MOD14 search one compound not identified within the 15 nearest matches is ethyl sorbate, even though the second closest match to this target is methyl sorbate. Similarly, with the MOD10 search on the target spectrum of 1-methyl-4-isopropylbenzene, the correct library entry is slightly farther from the target than the ortho and meta isomers and so appears lower in the list of nearest matches. Consideration of such effects, which are observed more frequently for search algorithms using compressed data, mitigates judgment against the somewhat poorer performance of these algorithms.

A similar graph summarizing the results of searches using the same five algorithms with the second set of target compounds is shown in Figure 3. As noted with the first target set, the search using complete mass spectral data performs best. The MOD14, MOD10, and binary searches all identify six targets with the single nearest match, but as more nearest matches are considered, the performances of the binary and MOD14 search remain comparable while the MOD10 performance falls off. Again, the performance of the MODN

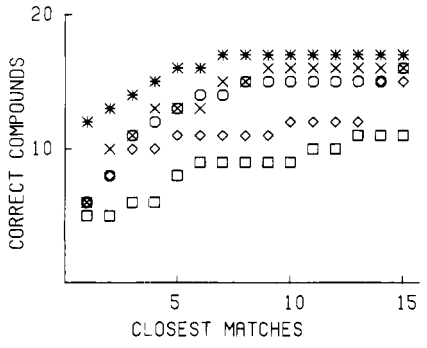

Figure 3. Results of searches on target set 2 for five data compression techniques: (*) uncompressed, (×) binary, (○) MOD14, (◇) MOD10, (□) MOD7.

searches declines as the degree of data compression increases. As anticipated because of the greater complexity of the target compounds, the performance of all searches is worse than that observed with the first target set. Ironically, complete mass spectral data is a disadvantage in two cases. The target riboflavin is recognized easily by the binary search but not by the search using complete data. Similarly, adenine is identified by the MOD14, MOD10, and binary searches but not by the search based on full mass spectra. In both cases, there are significant differences between library and target spectra which are masked by data compression. While this observation likely identifies an argument in support of careful data selection as much as it does an advantage of data compression, it is a factor that must be considered.

Two important points are illustrated by the trial searches with these sets of target compounds. One is that, although extensive data compression does hurt the performance of a search algorithm, the best algorithms using compressed data do approach closely the performance of algorithms using complete mass spectral information. Secondly, results obtained with the MOD14 search, which consistently shows the best performance of the MODN searches, are comparable to those obtained with the best alternative search algorithm with similar data storage requirements. These results indicate that MODN searches in general and the MOD14 search in particular can perform effectively as "mass spectral" search algorithms.

FILE PARTITIONING

One benefit that arises from the use of a MODN compression is the ability to reduce the time required for searches by limiting the number of library entries that need to be considered in detail by the comparison algorithm because the nearest matches to a target spectrum essentially always have the same most intense MODN dimension as the target. This implies that MODN spectra generally have a predominant ion series. Because the nearest matches to a given target are those spectra with the same most intense ion series, the search algorithm need compare only a target spectrum with those library entries having the same predominant ion series. At least two methods can be used to identify the library entries appropriate for comparison with a specific target spectrum. One approach is to include a number with each library entry that identifies the most intense ion series, but this method increases the size of the library file and still requires a brief examination of each library entry. Alternatively, all library entries with the same most intense ion series are grouped together in the data file. Each MODN spectrum then belongs to one of N partitions within the library file, and a set of file pointers identify the beginning and end of each partition. When a target spectrum is searched, the appropriate partition is selected according to the most intense ion series in the target spectrum, and only library entries in that partition are used for comparison.

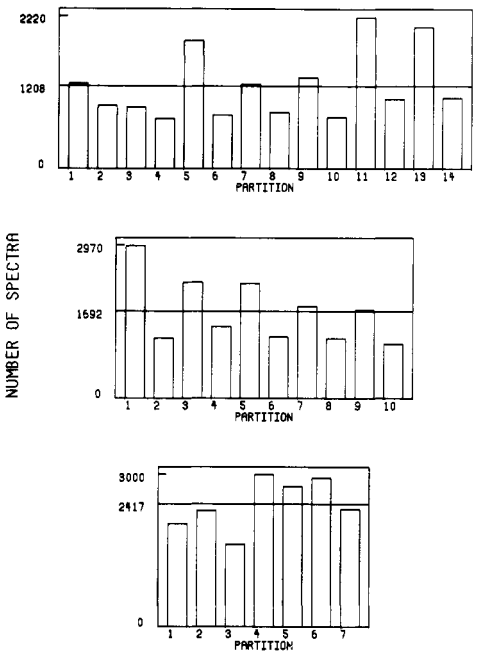

Figure 4. Distribution of library spectra per partition for three MODN compressions.

Table III. Partitioning Efficiency Calculations for Three MODN Compressed Library Files

	MOD14	MOD10	MOD7
observed partitioning information (bits)	3.70	3.24	2.78
maximum partitioning information (bits)	3.81	3.32	2.81
observed/maximum	0.97	0.98	0.99

One measure of the efficiency of the partitioning is suggested by information theory. The information associated with knowing which file partition to search can be estimated with eq 8, where p_i is the probability of a spectrum being in the

$$I = \sum_{i-1}^{N} -p_i \log_2 p_i \qquad (8)$$

ith partition, N is the number of partitions, and I is the information in bits.

Clearly the maximum information or efficiency is obtained if each partition is equally occupied. If all p_i's equal $1/N$, then knowing which partition to search reflects $\log_2 N$ bits of information. To evaluate the partitioning efficiency for the library used in this study, the fraction of the total number of patterns contained in each partition is taken as an estimate of the probabilities. Figure 4 shows the distribution of spectra in partitions according to the most intense ion series or dimension for each MODN compression. The solid line across each bar graph indicates the number of spectra that would be in a partition if each were equally occupied. Table III summarizes the results of calculations comparing the partitioning information observed for the library file with the information expected for a maximally efficient partitioning of the file. For this library, all MODN compressions approach closely the maximum possible partitioning efficiency.

An alternative indicator of the efficiency of the partitioning is to consider the reduction in the number of comparisons performed during a search. In a conventional search each target spectrum is compared with all library entries. With file partitioning each target is compared with only those entries in one partition. For a given set of target compounds the number of comparisons performed both with and without partitioning can be easily calculated. Furthermore, if the

Table IV. Comparison Ratios as Percentages Calculated from Target Set 1 and Library Statistics for Three MODN Compressions

	MOD14	MOD10	MOD7
from first target set	8.31	11.00	14.99
from library statistics	7.89	11.23	14.82

fraction of library entries in a partition is used as an estimate of the fraction of target spectra that will require a search of that partition, it is possible to estimate the savings directly from the partitioning statistics of the library file. The sum of the squares of the fraction of entries in each partition gives the ratio of the number comparisons required with file partitioning to the number comparisons required by the conventional search strategy. These ratios, as calculated from library statistics and for the set of 40 target compounds, are reported in Table IV. Because a search algorithm not only compares spectra but also performs tasks such as sorting new matches into a running list of nearest matches and looking up the names of nearest matches, the actual time required for a search of a partitioned file will not be simply the product of the time required for a conventional search and the fraction of the spectral comparisons performed.

It is also interesting to consider the relationship between MODN dimensionality and the use of file partitioning. To do this, the number of comparisons between spectra must be weighted to include the number of dimensions in a spectrum. The comparison of two MOD14 spectra requires the comparison of twice as many pairs of numbers as does the comparison of two MOD7 spectra. Thus, using the conventional search strategy a MOD7 search compares only half the number pairs that a MOD14 search does. However, with file partitioning, the number of spectra in a partition increases as the number of dimensions, and therefore partitions, decreases. MOD14 spectra are still twice as long as MOD7 spectra, but they are distributed among twice as many partitions. For the cases examined the partitioned MOD10 and MOD7 searches require consideration of about 95% as many number pairs as the partitioned MOD14 search. The use of file partitioning makes the search times required for typical MOD14 searches comparable to those required for the lower dimensionality MODN searches.

CONCLUSIONS

The use of MODN spectra offers a promising approach to the compression of mass spectral data for library searching. The MOD14 search probably represents the best choice of the three considering the trade-off between search performance and library storage requirements. The efficiency of the data compression, the molecular structure information implicitly contained in MOD14 spectra, the general performance of the search with target compounds, and the effectiveness of file partitioning are features which recommend use of this data compression method.

ACKNOWLEDGMENT

The authors wish to thank Alan Hanna for helpful discussions during the course of this work. Portions of the work were presented at the 29th Annual Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.

LITERATURE CITED

(1) "Index of Mass Spectral Data", AMD11, American Society for Testing and Materials, Philadelphia, Pa., 1969.
(2) S. Abrahamsson, *Sci. Tools*, **14**, 29 (1967).
(3) B. Petersson and R. Ryhage, *Ark. Kemi*, **26**, 293 (1966).
(4) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1464 (1968).
(5) H. S. Hertz, R. A. Hites, and K. Biemann, *Anal. Chem.*, **43**, 681 (1971).
(6) H. W. Brown and E. J. Bonelli, *Abstr. 1977 Pittsburgh Conf. Anal. Chem. Appl. Spectrom.*, 144 (1977).
(7) S. L. Grotch, *Anal. Chem.*, **42**, 1214 (1970).
(8) S. L. Grotch, *Anal. Chem.*, **45**, 2 (1973).
(9) B. A. Knock et al., *Anal. Chem.*, **42**, 1516 (1970).
(10) L. E. Wangen, W. S. Woodward, and T. L. Isenhour, *Anal. Chem.*, **43**, 1605 (1971).
(11) M. C. Hamming and R. D. Grigsby, *Proc. 15th Annu. Conf. Mass Spectrom. Allied Top.*, 107 (1967).
(12) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1469 (1968).
(13) D. H. Smith, *Anal. Chem.*, **44**, 536 (1972).
(14) S. P. Markey, W. G. Urban, and S. P. Levine, Eds., "Mass Spectra of Compounds of Biological Interest", USAEC Technical Information Center, Oak Ridge, Tenn., TID-26553-P1.
(15) R. M. Silverstein and G. C. Bassler, "Spectrometric Identification of Organic Compounds", Wiley, New York, 1967.
(16) S. L. Grotch, *Anal. Chem.*, **43**, 1362 (1971).

Computer-Assisted Examination of Chemical Compounds for Structural Similarities[1,2]

TOMAS H. VARKONY, YOSSI SHILOACH, and DENNIS H. SMITH*
Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305

Received August 17, 1978

An algorithm for finding common substructures among a potentially large and diverse set of chemical structures is described. The algorithm has been implemented in an interactive computer program called MAXSUB. The program allows the chemist to specify his definition of what constitutes "commonality" of substructures by providing control over the importance of degree of substitution, hybridization, atom type, and ring membership of atoms in substructures and multiplicity of bonds between atoms. Applications to problems involving topological representations of chemical structures, including macrolide antibiotics and marine sterols, are discussed briefly to illustrate the program. Some possible extensions to dealing with three-dimensional representations of structures are mentioned.

Comparison of structural features among a set of chemical structures which display some common behavior is a frequent problem in chemical research. This problem can usually be characterized as one of relating the structures of the molecules to some "activity", i.e., structure/activity relationships in the broadest sense of the term. For example, the activity may be of a physical nature in that the molecules all display some characteristic pattern or subpattern in a spectroscopic technique, or biological in that the molecules demonstrate similar physiological effects. The importance of relating common portions, or substructures, of the molecules to commonly observed activities, whether to build correlation tables of substructures to spectroscopic behavior or to design new drugs,[3] to mention only two applications, hardly needs emphasis. In all such studies it is presumed that molecules displaying similar activities do so because they share some similar feature or

0095-2338/79/1619-0104$01.00/0