EPA-600/4-76-046
October 1976

Environmental Monitoring Series

COMPUTER INTERPRETATION OF POLLUTANT MASS SPECTRA

Environmental Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Athens, Georgia 30601

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

EPA-600/4-76-046
October 1976

COMPUTER INTERPRETATION OF POLLUTANT MASS SPECTRA

by

Fred W. McLafferty
Cornell University
Ithaca, New York 14853

Grant No. R-801106

Project Officer

John M. McGuire
Environmental Research Laboratory
Athens, Georgia 30601

ENVIRONMENTAL RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
ATHENS, GEORGIA 30601

DISCLAIMER

This report has been reviewed by the Environmental Research Laboratory-Athens, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U. S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

FOREWORD

Nearly every phase of environmental protection depends on a capability to identify and measure chemical pollutants in the environment. The Analytical Chemistry Branch of the Athens Environmental Research Laboratory develops techniques for identifying and measuring chemical pollutants in water and soil.

This report describes two computer programs that assist chemists in identifying organic compounds from their mass spectra. One program rapidly selects, from a computer file, spectra that have a high probability of matching the spectrum of an unidentified compound. The other program greatly assists the analyst in postulating the identity of a compound whose spectrum is not in the file. The programs will significantly enhance the assessment of health and ecological effects of organic chemicals and the development and implementation of control measures.

ABSTRACT

The objective of this research was to improve systems for computer examination of the mass spectra of unknown pollutants. For this we have developed a new probability based matching (PBM) system for the retrieval of mass spectra from a large data base, and have substantially improved the interpretation of unknown mass spectra using the self-training interpretive and retrieval system (STIRS). PBM was designed as a prefilter to STIRS; if an unknown mass spectrum can be identified with a sufficiently high confidence by PBM, interpretation of the spectrum using STIRS is not necessary. The PBM system provides more efficient retrieval than presently accepted systems; it incorporates a "reverse search" algorithm, and through the use of weighted mass and abundance data provides a statistically valid prediction of the confidence of the matches found. STIRS has been improved to give a confidence-level prediction of the presence of ~200 particular substructural features in the unknown molecule. Extensive studies have been made to improve the data selection for most data classes used by STIRS, resulting in a much higher level of overall system performance. Operation efficiencies of both PBM and STIRS have been improved dramatically so that both require less than 1 minute on a laboratory PDP-11/45 computer. STIRS has been made available for outside use by long-distance phone connections to this PDP-11/45, and recently both PBM and STIRS have been made operational on the Cornell IBM 370/168 so that these are available internationally over the TYMNET computer network system.

TABLE OF CONTENTS

Sections

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author is pleased to acknowledge the very substantial research contributions of Drs. Rengachari Venkataraghavan, Gail M. Pesyna, and Henry E. Dayringer, which made possible the results described here, and the very stimulating suggestions and interactions with Dr. John M. McGuire.

SECTION I

INTRODUCTION

IDENTIFICATION OF POLLUTANT MASS SPECTRA

The overall objective of this project was to develop a system for the automatic identification of pollutant molecules utilizing a gas chromatograph/ mass spectrometer (GC/MS). The original objective was to improve the Cornell self-training interpretive and retrieval system (STIRS) for automated computer identification and interpretation of mass spectra through extensive testing, statistical evaluation of results, modifications for optimized performance, and evaluation of the effect of data quality. During the project period, STIRS was used extensively by other United States scientists through a phone-line connection to our laboratory PDP-11/45 computer developed under this grant. This provided dramatic evidence that a large proportion of users primarily needed a matching system; in a substantial proportion of cases, their unknown spectrum did not need to be interpreted because a reference spectrum of the same compound was already in the file. Thus we decided to develop a matching system as a prescreen to STIRS so that the latter would only be used for those unknown mass spectra for which a sufficiently good match could not be found in the reference file. The two main sections of this report therefore concern the development of PBM and the improvement of STIRS. We have recently prepared an extensive review of this field entitled Computerized Structure Retrieval and Interpretation of Mass Spectra,[1] to which the reader is referred for more details on the problems involved and previous computer system proposals.

PBM

In designing PBM, we attempted to utilize principles employed by information retrieval scientists in systems developed for such problems as document retrieval from libraries.[2] We thus incorporated a reverse search concept,[3,4] which is especially valuable for unknown spectra of mixtures and "weighting" of the mass and abundance data used. Determination of these weighting factors was based on the statistical occurrence of the mass and abundance values found on examination of a large comprehensive data base of 18,806 spectra.[5] The PBM concept was initially developed with R. H. Hertel and R. D. Villwock[3] for use with a quadrupole mass spectrometer operated under feedback control by a dedicated microcomputer. However, this system utilized a data base of only 16 or 64 mass spectra, all of which had been determined on the same instrument. Thus a major problem in the design of the more general purpose PBM system involved the fact that the reference mass spectra had been determined on a wide variety of instruments under diverse experimental conditions, and that the spectra contained many artifacts such as impurity peaks and incorrect mass and abundance data. The solution to this problem was to have the computer repeat the comparison calculation a number of times, each time dropping another selected peak of the reference spectrum (the so-called "flagged peaks") and retaining only the highest confidence value achieved. Thus a few impurity or aberrant peaks would not eliminate the spectrum as a possible match.

Statistical Basis for PBM

Probability based matching is based on the general rule of multiplication of probability theory[7] which states that if n independent events occur with probabilities p_1, p_2, ... p_n, then the probability of all n of these events occurring is given by equation 1. Thus if peaks with mass m_1 and m_2 having intensities i_1 and i_2 occurring in mass spectra with probabilities p_1 and p_2, the probability that both occur at random in an unknown spectrum is $p_1 \times p_2$. If this product is small, it is much more likely that the presence of peaks m_1 and m_2 in intensities i_1 and i_2 is due to the identity of the unknown spectrum with that of the reference compound.

$$\text{overall probability} = \prod_{i=1}^{n} p_1 \qquad (1)$$

The low value of this probability provides a confidence that an identification is correct, measured by a "confidence value" or "K value." This measure, as well as all the individual probabilities, is expressed as the corresponding base two logarithm for convenience of calculation; inverse probabilities are also used to simplify the calculations and to produce a final result that is a direct measure of "confidence." In this reverse search, there is computed for each reference spectrum matched against the unknown a confidence value, K, equal to the sum of the individual K_j values. K_j is calculated for each peak in the unknown whose intensity agrees within a predetermined range to that of the corresponding peak in the reference spectrum. K_j combines four terms,

$$K_j = U_j + A_j + W_j - D$$

where U_j is the contribution to the probability of the "uniqueness" of the m/e value of peak j; A_j is the contribution to the probability of the abundance value of the peak as it appears in the reference spectrum; W_j, the "window factor," is a measure of the agreement required between the abundance of the peak in the reference and in the unknown; and D, the "dilution factor" for mixture spectra, is a measure of the overall reduction of peak intensities in the unknown due to the presence of other components (if the unknown spectrum is of a pure compound, D = 0).

A peak in the unknown that does not agree within the window tolerance is ignored in the cumulative calculation of confidence. If it is more intense than would be expected, it is termed "contaminated." However, peaks of intensities less than the minimum allowed are treated differently than in the earlier PBM system. In that system, all the reference and the unknown spectra were recorded on the same instrument, and the background level was known for each unknown sample; therefore it was assumed that a peak in the unknown could not be of lower relative intensity if that reference compound was present in the unknown. In the present system, on the other hand, which uses a large data base of spectra from diverse sources, this could be true because of experimental variation or even impurities in the reference compound; thus a limited number of less intense peaks are "flagged" in the match to ignore this discrepancy.

The assumption that mass spectral peaks are independent events, which is essential to the rigorous application of the general rule of multiplication, is of course far from exact for many mass spectral peaks. For example, it is much more common to find $\underline{m}/\underline{e}$ 41 in a spectrum containing an abundant $\underline{m}/\underline{e}$ 57 peak. It would also be expected that the molecular ion and other high mass peaks would show less cross-correlation, and so these are given extra preference.

STIRS

The first major improvement was also suggested by the phone-line operation. Information on the type of molecule giving the unknown and mass spectrum is obtained by examining the reference compounds selected by STIRS to be the most closely related to see if these contain common structural features; thus if 10 of the top 15 compounds contained an imidazole ring, this would indicate with relatively high probability that the unknown also contained imidazole. Obviously the significance of this observation would be much greater if only 0.1 percent of the reference compounds in the file contained imidazole than if, for example, 10 percent of them did. Thus for this improvement the computer examines the top selected compounds for the presence of a variety of selected substructures, and calculates the probability that the number found could be due to a random selection instead of to the actual occurrence of the substructure in the unknown molecule.

The successful development of such a quantitative evaluation procedure for STIRS performance with a wide variety of substructures then also provides a method to see if STIRS modifications actually improve performance. In the basic operation of STIRS, the mass spectral data of the unknown of a particular type is compared against the corresponding data of all the reference mass spectra to find those compounds of best match. The original selection of these data classes was based on mass spectrometry knowledge and intuition, plus relatively qualitative performance tests. The substructure quantitative results thus allow the evaluation of changes in the selection of data class according to the number of peaks, whether they are odd- or even-electron ions, the extent of the mass range, overlapping of mass ranges, various combinations of data classes, and special series of mass spectral peaks. The ranges are incorporated to optimize the "recall", the proportion of compounds examined which actually contain the particular substructure for which STIRS was able to identify that substructure with greater than 98 percent confidence.

Basic Principles of STIRS

A number of classes of mass spectral data known to have high structural significance, such as characteristic ions, series of ions, and masses of neutrals lost, are identified; for each class the computer matches the data of the unknown mass spectrum against the corresponding data of all reference spectra. In each data class the reference compounds whose spectra have the highest "match factor" (MF) values are examined by the chemist for any common structural features, with a high frequency of occurrence indicating a high probability that the structural feature is present in the unknown.

Substructure Identification

The 15 selected compounds of highest MF value are examined by the computer for the presence of specific substructural groups to provide a statistical evaluation of the probability of the presence of each group in the unknown compound. The principle used for this statistical evaluation is random event theory.[8,9] In a manner similar to calculation of the odds in random drawing from a collection of colored balls, the probability of selecting a compound at random containing any given substructure can be calculated knowing only the percentage in the file of compounds that contain such a substructural unit. The probability P that any particular number N of compounds containing a given substructure out of 15 compounds has been drawn at random is

$$P(N) = 15!(x)^{N}\ (1 - x)^{(15\ -\ N)}/[N!(15 - N)!]$$

where x is the decimal fraction of the file compounds with the substructure. To evaluate the importance of this probability it is compared to that of the selection of the most probable number. (This is a more conservative estimate than comparing it to all more probable events; the difference is not large, especially for substructures of infrequent occurrence, and the precision of this estimate is also evaluated experimentally.) The most probable number of compounds containing the specific substructure to be drawn out of 15 is $15 \cdot x$, so the relative frequency of finding N compounds in the top 15 is $P(N)/P(15 \cdot x)$. A ratio of 1/200 predicts that on a random basis STIRS would retrieve the most probable number $(15 \cdot x)$ of compounds with the substructure 200 times (on average) before it retrieves N compounds with the substructural unit. Thus if STIRS does retrieve N compounds with the substructure, only 1 time in 200 will this be due to chance, so that this result gives a 99.5% confidence that the substructure is actually in the unknown.

Consider as an example an unknown compound analyzed by STIRS match factor 11 for the presence of the phenyl group. Based on the fact that 28.4% of the compounds in the reference file contain phenyl, if the unknown does not contain phenyl an average of 4.2 phenyl-containing compounds would still be expected in the top 15 compounds selected as matches in MF11. If 10 of the top 15 compounds actually contain phenyl, the probability that this occurred at random is $P(10)/P(4.2) = 1/113$; that is, this result indicates with >99% confidence that phenyl is in the unknown.

To evaluate the method, it has been applied first to 204 substructures commonly found in organic compounds. These were chosen to test STIRS' ability to identify a broad range of structural features and would be expected to vary widely in their mass spectral behavior. For example, the carbonyl group, in contrast to its characteristic effect on the infrared spectrum, and despite its strong directing effect on mass spectral decompositions, does not produce peaks of masses unique to itself. Even in the case of functional groups which do give characteristic peaks, such as the terminal benzoyl group's ions at _m/e_ 105 and 77, the abundance of such peaks can be greatly reduced by the presence in the molecule of an additional functional group which directs the fragmentation more strongly, such as _p_-$NH_2C_6H_4CO$-. Thus even if the presence of a particular ion (or ions) in the mass spectrum is reliable evidence of the presence of a particular substructure, _the absence of that ion does not necessarily show that the group is absent_. Thus, although most mass spectral learning machine methods[18-21] are designed to give "yes/no" answers on the presence of structural features, we have not attempted to obtain negative information from STIRS, restricting our consideration of STIRS substructure predictions to those of ≥98% confidence. To evaluate these predictions, a large spectral collection[6,17] has been used to gather statistical data on the "information precision" and "recall" of the results. The definitions of these terms are patterned after those of terms used by information scientists to evaluate the efficiency of, for example, a document retrieval system. Because the term "precision" has a somewhat different meaning to chemists, we will use the modified term "information precision" to mean the percentage of substructure predictions which are actually correct. The "recall" value is the percentage of compounds actually containing the substructural group for which the group is identified by STIRS. A more detailed discussion of the usefulness of such precision and recall values in evaluating mass spectral analysis systems will be presented separately.

Data Class Improvements

To improve the frequency of correct answers (recall) and reliability of answers (information precision) for the extraction of substructural features by STIRS, we have tested the following four approaches, embodied in the characteristic ion data classes shown in Table 1: (1) Variation of the number of ions used in matching for each mass range; (2) Forcing the use of both even- and odd-electron ions instead of using the largest peaks regardless of mass; (3) Use of additional mass ranges overlapping the original set (for example, beside _m/e_ 6 - 88, MF2A, and _m/e_ 89 - 158, MF3B, including _m/e_ 61 - 116, MF3A); (4) Arithmetic combinations of the MF values found by several data classes (this approach was previously shown[10] to be useful in the "overall match factor", MF11, a combination of MF1 through MF6).

TABLE 1. TESTED DATA CLASSES FOR CHARACTERISTIC IONS

Data class	Maximum number of peaks	Mass range
2[a,b]	3 even-mass, 3 odd-mass	6 - 89
3[b]	5	90 - 149
4[b]	5	150 - $(M - 1)^+$
2a[c]	5 even-mass, 5 odd-mass	6 - 89
2a'[c]	10	6 - 89
2A	4 even-mass, 4 odd-mass	6 - 88
2A'	8	6 - 88
3A	8	61 - 116
3B	8	89 - 158
4A	8	117 - 200
4B	8 (4, 6)[d]	159 - 270
4C	8 (5)[d]	201 - $(M - 1)^+$
11.1	2A + 3A + 3B + 4A + 4B + 4C	
11.2	2A' + 3A + 3B + 4A + 4B + 4C	

[a] The degree to which the class 2 data of the unknown spectrum match those of the reference is given by STIRS as the "match factor 2 (MF2)" value.

[b] Used in the original STIRS program (ref 10).

[c] This data class was tested with only 20 of the most commonly occurring substructures.

[d] Tests were also made with alternative maximum number(s) of ions.

SECTION II

CONCLUSIONS

Even in their present stage of development, we believe that PBM and STIRS constitute the most powerful combination system available for the computer examination of unknown mass spectra. PBM is the most rapid and efficient system for matching an unknown spectrum against a large and diverse reference file. This system should be especially useful for routine identification of unknown pollutants because its "reverse search" is uniquely appreciable to mixtures, and because it provides a confidence level measurement of the probability that the match is correct, utilizing several "match classes" of the degree of structural similarity of the unknown and reference compound. For those unknown spectra that cannot be matched with this sufficiently high degree of confidence by PBM, STIRS can often provide partial or even complete information on the molecular structure with a direct evaluation of the confidence. Because STIRS is meant to be an aid to, not a replacement for, the trained mass spectrometrist, the best answer in difficult cases will be achieved through human interpretation of computer results. In many instances, STIRS provides structural information that was not discerned by the trained mass spectrometrist, and the speed with which STIRS information with quantitative probabilities can now be obtained should mean that its rountine use would provide substantially increased confidence and efficiency in the results of human interpretation.

SECTION III

RECOMMENDATIONS

It is strongly recommended that the combined PBM/STIRS system be used and evaluated extensively on real problems by EPA mass spectrometrists. A few people should make a sincere effort to apply PBM and STIRS to every unknown spectrum encountered over, for example, a 1-month period. Their feedback to us would be invaluable for further improvements, and these key people could also train others in the future. We feel strongly that there still is a very real education problem in mass spectrometry. The author still gives a basic course on interpretation of mass spectra to hundreds each year, including many from EPA laboratories. Education in the basic principles and in new developments such as computer examination of mass spectra is a continuing problem; although this is not unique to mass spectrometry, it deserves the careful attention of all parties in the field.

Further improvements should be valuable for both the PBM and STIRS systems. "Real-time PBM" should be incorporated directly into the GC/MS/computer systems now used routinely for pollutant identification in major EPA labs. A possible system for this would involve computer collection and reduction of the GC eluent mass spectrum every 2 seconds; PBM matching of the most recently acquired spectrum against a data base would be run by the computer as a background operation during the next 2-second mass spectral collection. With proper GC/MS calibration, the computer could even calculate the quantity of identified components. Thus for complex pollutant samples, the GC/MS/computer system could give direct identification and quantitation of many components, greatly reducing the burden on the operator and mass spectrometrist.

SECTION IV

EQUIPMENT AND REFERENCE MATERIAL

Almost all of the experimental work was carried out on a Digital Equipment Corporation PDP-11/45 and -11/10 dual processor configuration with 28K and 24K of core memory, respectively, connected with a special bus window to make all peripherals addressable by either processor, removable disks of 1.2 M, 1.2 M and 29 M word storage capacity, DEC-GT/40 cathode ray tube display, printer, plotter, 9-track IBM-compatible magnetic tape drive, and telephone modem link for dial-up use by outside investigators. Most programs were in assembler language. Both PBM and STIRS performance evaluations were made by running approximately 400 unknown mass spectra for each item under investigation which often required continuous runs of many days to achieve statistically meaningful data. Evaluation of the STIRS system by outside users was done through a modem and interface for telephone processing of the submitted unknown spectrum. Users were contacted by both phone and letter.

DATA BASE

The Registry of Mass Spectral Data, representing 18,806 different compounds[6] was used for STIRS, and these data plus 5,073 lesser-quality spectra of some of those compounds were used in the creation of the PBM library. The most recent work has been done with an expanded data base of >35,000 mass spectra. Although a large number of errors had been eliminated during the original preparation of the data base, checking of individual cases of poor results showed that a significant number remained.

PBM U AND A VALUES

The Registry file of mass spectra of 18,806 different compounds was used to determine the probabilities of occurrence of the mass values of peaks of ≥1% abundance.[5] Although the uniqueness of peaks fluctuates substantially for *m/e* 29 - 114, as expected, there is a linear decrease in occurrence probability >*m/e* 114, being reduced by half approximately every 58 mass units. A data base of much higher molecular weight should be reduced linearly by a substantially smaller factor, approaching a halving every 130 mass units; these probabilities were used for the PBM "U" values. The abundance values for masses >*m/e* 120 were found to show a surprisingly constant distribution which is log normal; for lower abundance values this distribution is dependent on and predictable from the molecular weight range. The resulting data were used for the PBM "A" values.

PBM CONDENSED REFERENCE FILE

The following metastable, multiply charged and impurity peaks are eliminated from each reference spectrum before it is condensed: all peaks having non-integral masses, peaks at *m/e* 18, 28 and 32 which may be due to water and air, and peaks found at masses higher than those in the molecular ion cluster, with the limit defined as the molecular weight + 3 + 2(# of Cl atoms + # of Br atoms) + 1/2(# of S atoms + # of Si atoms). If the compound does not contain elements other than the most common ones, and if its molecular weight is greater than 50 amu, peaks due to the illogical neutral losses of 4 - 12 amu and 21 - 23 amu are also excluded, plus the loss of 18 amu if the compound does not contain oxygen, the loss of 19 amu if no oxygen or fluorine, the loss of 20 if no fluorine, and the losses of 13, 14, 24 and 25 amu if no chlorine or bromine (in the latter case these may be isotopic ions of losses of 15, 16, 26 or 27).

The reference spectrum is renormalized if necessary and the U value of each peak is determined. All peaks below mass 29 are arbitrarily assigned U values of 1, a value which is low enough to actively discriminate against the selection of these peaks but still permits them to be used if the spectrum contains no peaks or very few peaks at higher *m/e* values.

The peak abundance percentages have been divided into standard ranges assigned to specific A values.[5] For the reference spectrum the A value of each peak is determined by the range into which its abundance falls. Thirty-two of the amu values have abundance probability distributions significantly different from the standard distribution, so that special abundance ranges must be used for these A value determinations.

All peaks in the spectrum are ordered by decreasing U + A values; within each set of peaks having the same value of U + A, the peaks are ordered on the basis of decreasing *m/e* values. The 15 peaks at the top of this ordered list are checked for the presence of the base peak, for the most abundant isotopic peak in the molecular ion ($M^{+\cdot}$) cluster, and for the peak (or two peaks if $M^{+\cdot}$ is not present) corresponding to the neutral loss(es) of 18, 20, 27, 28, 30, 32, 34, 36, 42, 44, 46, 48, 56, 60 or 64 amu having the largest U + A value(s). If any of these three are not already included in the list, they are substituted for the peaks of lowest (U + A) value.

For each reference spectrum the serial number, lowest recorded mass in the spectrum, $M^{+\cdot}$, and the values of *m/e*, abundance, and U + A for the 15 selected peaks are packed into 32 computer words and stored in a file which occupies 1494 blocks of 512 16-bit words each of disk storage. The disk file structure is optimized to reduce access time.

STIRS CONDENSED REFERENCE FILE

The STIRS file data were prepared as described earlier,[10] except for the information on substructures and new match factors. The WLN notation of each compound was used to assign a computer bit fragment code for each of the 204 substructures examined; the linear notation of WLN is particularly appropriate for such mass spectral relationships because its units often correspond to the pieces of the molecule giving rise to spectral peaks. Note that only abbreviated definitions are given in the Tables; for example, the class "U" includes most double and triple bonds, but not carbonyl (V) and phenyl (R) groups. These groups were taken mainly from the "Dictionary of Frequently Found Substructures";[22] possible alternative WLN notations (for example, "2O" as well as "O2" indicates ethoxy) have been tabulated.[16,17] For each MF data class a count of the number of occurrences of each of the 204 substructures was made using the pre-generated fragment bit file to give the respective values of the decimal fraction x. This was used to calculate the value of N, the number of compounds of the top 15 containing the particular substructure necessary to give $P(N)/P(15 \cdot x)$ ratios of 1/50, 1/200, 1/500, and 1/1000 (confidence levels of 98%, 99.5%, 99.8%, and 99.9%).

For the new data classes MF2A-4C, the mass ranges chosen were based on the frequency of occurrence or "uniqueness" of mass values,[5] the limits of the ranges being selected to correspond to masses of high uniqueness. Because the occurrence frequency decreases at higher masses, the width of the ranges was increased for data classes of higher mass.

SECTION V

EXPERIMENTAL AND EVALUATION PHASE

PBM

System Design

Unique features of PBM include "reverse search"[3,4] and the "weighting" of the *m/e* and abundance values of the mass spectral peaks. The system was designed in particular to emphasize high information precision and retrieval, as those unknowns for which only low confidence matches can be achieved usually must also be examined by a mass spectrometrist or an interpretive algorithm such as STIRS.

PBM Search Algorithm

For each reference spectrum the search algorithm begins by examining the unknown for the presence of the reference peaks from highest to lowest *m/e* values. If a peak is not present it is flagged; if the number of missing peaks exceeds the number of allowed flagged peaks the program proceeds to the next reference spectrum.

If reference peak j is found in the unknown ($j = 1, 2, .., 15$), the ratio, ρ_j, of its abundance in the unknown to its abundance in the reference is calculated, and if ρ_j is less than the specified "minimum percent component" (which for these studies was 10% for pure compounds and 1% for mixtures unless otherwise specified), peak j is flagged; ρ_j values are calculated for all such reference peaks unless the maximum number of flagged peaks is exceeded. The smallest ρ_j *not* associated with a flagged peak (ρ_{min}) is determined, and the confidence value K_j is calculated for this peak. ρ_{min} specifies the smallest percentage of this reference compound which could be present in the unknown sample and thus directly determines the dilution factor, D. The product of the abundance of each reference peak and ρ_{min} is the abundance expected for that peak in the unknown spectrum. The reference abundance also determines the window tolerance that is demanded of the match. For these studies a ±30% tolerance was permitted for peaks of abundances $\geq 9\%$, ±39% for 3.4 - 9% peaks, ±46% for 1 - 3.4% peaks, ±51% for 0.24 - 1% peaks and ±71% for peaks less than 0.24%; this gives eight "abundance bins," so W = 3.[5] The expected abundance of the unknown peak is set at the bottom of the ±*x*% window and the top of the window is determined. If the actual abundance of the unknown peak falls within these limits, K_j is calculated for it from $U_j + A_j$ (determined by the peak in the *reference* spectrum) $- D + W$, and added to the accumulated K value. If the abundance of the peak in the unknown spectrum is higher than the top of this window, it is termed "contaminated," and $K_i = 0$. (From the definition of ρ_{min}, the abundance of the peak will never fall below the window). After the entire set of reference peaks has been examined, one factor of W is subtracted from the K value because the peak which gave rise to ρ_{min} is guaranteed to fall within the window.

The K value resulting from the match is compared with the threshold K value (an optional threshold, 25 being used in this study); if K is smaller than the threshold it is not stored as one of the results. Otherwise the "percent contamination" is calculated, and if it does not exceed a specified maximum, the K value is stored. The "percent contamination", which is only an estimate of the true value, is calculated using the ten peaks of highest U + A in the unknown and is based on the proportion of the peak abundances which fall above the predicted abundance windows. If a maximum percent contamination less than 100% (i.e., none of the 10 unknown peaks are contained in the reference spectrum) is specified, reference compounds of molecular weights less than the masses of any of the ten peaks are not examined. Unless otherwise specified, for this study the maximum percent contamination was set at 20% for pure compounds and 100% for mixtures.

If the number of flagged peaks allowed has not been reached, the peak which gave rise to ρ_{min} is flagged and dropped from consideration, the next lowest value of ρ_{min} substituted, and the matching algorithm is executed again. This determines a new K value for this reference spectrum which is stored in place of the previous value if it is higher. When no more flagged peaks are allowed, the next reference spectrum is examined. In the studies reported here a maximum number of three flagged peaks was allowed for pure compounds and two for mixtures unless stated otherwise.

With each K value reported, a ΔK value is also calculated and displayed. The ΔK value is the difference between the K value found and the maximum value that could have been achieved by a perfect match with the reference spectrum.

Methods for Evaluating PBM Performance

A "low molecular weight set" (LMWS, MW 144-160 amu) and a "high molecular weight set" (HMWS, MW 232-312 amu) of unknown spectra were created for testing PBM's performance on the spectra of both pure compounds and mixtures. The sets of "pure" spectra were composed of 433 and 415 spectra (LMWS and HMWS, respectively) which are other spectra of compounds represented in the 18,806 spectra of the data base. These test spectra represent all those available in the "duplicate" portion (serial numbers 18,807 - 23,879) in the Wiley magnetic tape[6] except that a small number of spectra of impure and isotopically labeled compounds were excluded. A spectrum in the "Registry" file which is of one of these compounds was combined with two others in the ratio of 60:30:10 to create sets of synthetic mixture spectra containing 102 and 80 spectra, respectively, in the LMWS and HMWS.

To analyze the results obtained by the retrieval system, two parameters taken from the field of document retrieval[2] are defined: the recall, or the proportion of all possible matches which are actually retrieved, and the information precision, or the proportion of the retrieved spectra which are matches. These recall-precision pairs are computed at various retrieval levels: e.g., at particular K value levels. The trade-off between recall and precision is evident when recall is plotted on the x-axis, precision on the y-axis, of a plot such as that in Figure 1. An ideal retrieval system would maintain a precision of 1.0 at every level of recall: i.e., all matches and no mismatches are retrieved with high K values. A system whose recall-precision plot more closely approaches this ideal plot is the better system.

The precision achieved will depend on the degree to which the structure of the retrieved reference spectrum is required to match that of the unknown; for example, mass spectra are not sensitive to optical isomerism, and so such a requirement should not be imposed. These studies used the four somewhat arbitrary classifications shown in Table 2.

The molecular ion is uniquely characteristic, being especially valuable for differentiating the spectra of homologs. Thus a retrieved compound whose molecular ion is present and used in matching is given a special "+" notation. For these the degree of match is designated by "K+" and "ΔK+" values and the match classes by I+ through IV+. The K+ recall values shown are based only on the number of possible matches which contained a molecular ion; this represented 82.8% of the total possible matches for the HMWS.

To compare the performance of the PBM system with other retrieval systems, each spectrum in the file of 23,879 spectra was abbreviated using the two largest peaks in every 14 amu interval from m/e 6 to the highest recorded mass value, and the retrieval system designed by Biemann et al.[11] was implemented on the PDP-11/45 computer. The only difference between the system tested here and that described[11] is the elimination of the prefilter which specifies that the total abundance of homologous series of ions must be similar in the unknown and reference spectra; this should only result in a slower search and should not appreciably lower (it could improve) the matching capabilities. Exactly the same trial "unknown" mass spectra were tested on this and the PBM systems.

Methods for Evaluating STIRS Performance

The search algorithm for STIRS has already been described in detail,[10] and the computer program for STIRS is available through DECUS, Maynard, Massachusetts 01754. Testing of the validity of the random drawing model for STIRS operation was performed by running randomly selected sets of spectra and calculating the individual probabilities for the presence of all 204 substructures based on the 15 highest MF value compounds selected for each data class. The compounds used as unknowns for a particular substructure were selected at random by taking every fiftieth compound in the file starting at number 125, a total of 373 spectra. If this gave less than 30 compounds containing the substructure, additional spectra of such compounds were selected at random to give 30 (or the total number in the

TABLE 2. ARBITRARY CATEGORIES FOR DEGREE OF STRUCTURAL MATCH

Class of match	Relationship of reference and "unknown" structures	Example of ref. compounds matching *cis*-1,4-dimethylcyclohexane as the unknown
I	Identical compound or stereoisomer	*trans*-1,4-dimethylcyclohexane
II	Class I or ring positional isomer	1,3-dimethylcyclohexane
III	Class II or a homolog	diethylcyclohexane
IV	Class III or an isomer of a class III compound formed by moving only one carbon atom	trimethylcyclopentane

file if smaller). For 20 of the more abundant substructures STIRS runs were made on two other sets of 373 spectra starting at numbers 130 and 135; the results from the three data sets were the same within experimental error.

A tally of the number of correct and incorrect answers at the ≥98, ≥99.5, ≥99.8, and ≥99.9% confidence levels were obtained for each MF and substructural group; only those groups predicted by STIRS as being present with a confidence above the required level were examined to see if they were actually present in the unknown compound. The recall value is the percentage of compounds containing the substructure in which it was correctly identified. The precision should reflect the occurrence of correct answers relative to the total of those correct and wrong; because only positive answers are considered (vide supra), only compounds containing the substructure can give correct answers, and only compounds which do not contain the substructure can give wrong answers. To compensate for the difference in the number of possible correct and wrong answers in the spectra examined, the precision was calculated as the recall value divided by the sum of this value and the percentage of compounds not containing the substructure in which STIRS indicated its presence.

$$\text{Precision, \%} = \frac{\dfrac{\text{number right answers}}{\text{possible right answers}}}{\dfrac{\text{number right answers}}{\text{possible right answers}} + \dfrac{\text{number wrong answers}}{\text{possible wrong answers}}} \times 100$$

Thus for a case in which 70 of the 373 tested compounds contained the substructure, and STIRS indicated that 35 of these 70 and 3 of the remaining 303 compounds contained the substructure, the recall would be 35/70 or 50% and the precision 0.5/(0.5 + 3/303) or 98%.

SECTION VI

DISCUSSION

PERFORMANCE OF THE PBM SYSTEM

A discussion of the PBM system in much greater detail can be found in the Ph.D. thesis of Dr. G. M. Pesyna.[12] The general behavior of the PBM system is shown by Figure 1; note that for these tests the spectrum of the "unknown" was not present in the data base, because the resulting "perfect match" would give an unrealistic bias to the evaluation. As found for other retrieval systems,[2] increasing the matching criteria of K or ΔK values increases the precision of the results but decreases the proportion of unknown spectra for which matching compounds can be recalled. The results of the LMWS and HMWS are qualitatively similar; the poorer performance with increasing molecular weight is consistent with the increasing number of possible compounds of a given molecular weight which must be differentiated using the same number of peaks.

The precision achieved can be compared to the values of K on the basis of the original definition[3] of the "confidence value" as the base 2 logarithm of the number of spectra of other compounds which would have to be selected at random in order to find one which matches the unknown spectrum as well as does the reference spectrum in question. Because the reference file used[6] contains approximately two^{15} spectra, a matching criterion of $K \geq 15$ should select only one wrong answer on a random basis. If there is one correct answer in the file which is also retrieved, this would correspond to a precision of 50%; in the same way a criterion of $K \geq 20$ should select a wrong answer only one time in 32, or ~97% precision. Obviously much higher K values are actually required to attain such precision values (Figure 1). In substantial part this discrepancy results from the recognized[3] inadequacy of the original assumption that the statistical uniqueness of a particular peak is independent of the presence of other peaks in the spectrum. It is not surprising that particular combinations of masses, such as the "ion series", occur much more frequently than predicted from equation 1 using the probabilities of the individual masses. This will be illustrated below in the discussion of classes of matches.

It should also be noted that the precision values would be substantially higher if only the reference spectrum matched with the highest K value had been considered as the answer (instead of all spectra of K greater than a particular value), which is the more probable way that PBM will be used in practice. Thus if there are actually several reference spectra of the same compound as the unknown, this would substantially increase the probability that one of these reference spectra would be found to have the highest K

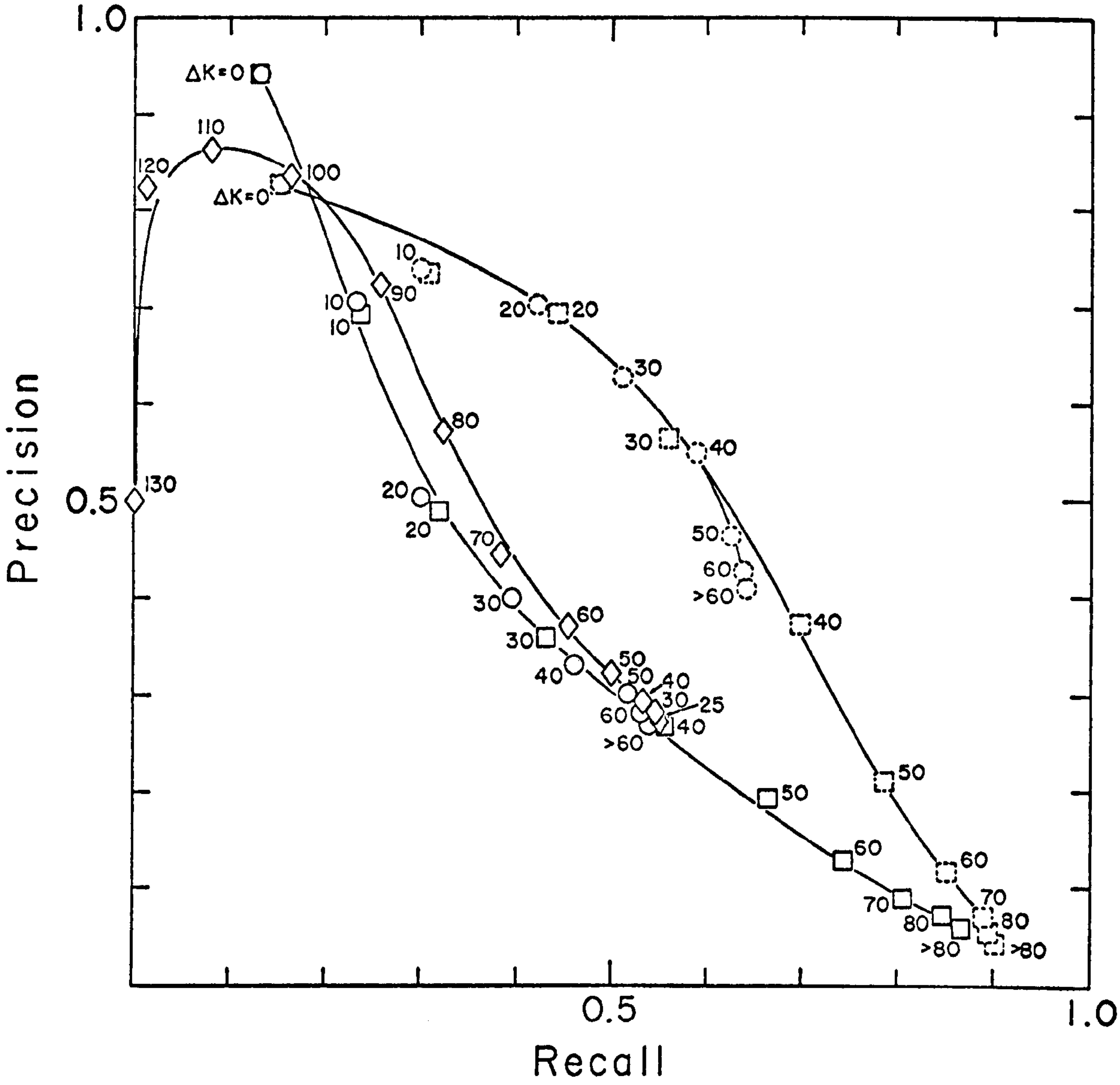

Figure 1. PBM performance for unknown mass spectra of pure compounds. LMWS, ΔK values: ◌, maximum percent contamination (MPC) = 20; ⬚, MPC = 70. HMWS, ΔK values: ○, MPC = 20, and □, MPC = 70. HMWS, K values: ◇, MPC = 20.

value. Of course much higher precision would be found using only reference spectra obtained under the same experimental conditions as used for the unknown, but limitation to such a file would be a serious handicap for a completely unknown spectrum.

Number of Flagged Peaks

Recalculation of the degree of match ignoring the peak of lowest abundance (unknown spectrum relative to the reference) was done to avoid mismatches due to impurities in the reference and other errors; the data for the LMWS (Figure 2) and HMWS[12] show that this is indeed beneficial. Increasing the number of flagged peaks from zero to two increases the maximum recall for LMWS from 40% to 61%, and for HMWS from 34% to 50%. The increase with the third flagged peak is much smaller, and the fourth has a nearly negligible effect. Because the flagged peak calculations increase the time requirement, three flagged peaks appear to be an optimum number for the present system.

Under the constraints used to produce the data of Figure 2, the recall increase which can be achieved by lowering the matching requirements reaches a limit. Incremental increases in the required ΔK value become less effective in increasing recall, approaching a maximum at ΔK or ~50. This is because most pairs of the unknown and reference spectra of the same compounds that could not be matched at that tolerance level contain a particular datum which eliminates the reference spectrum from further consideration. The increase in the maximum recall value (Figure 2) with increasing number of flagged peaks shows that this technique compensates for part of these incompatible data.

Increasing the tolerance to incompatible data by "flagging" that peak also increases the probability that the spectrum of an incorrect compound will be selected as a match; note (Figure 2) that the precision corresponding to a ΔK of 40 drops from 0.69 using zero flagged peaks to 0.51 using four flagged peaks. For the HMWS the precision values achieved at particular recall values are also somewhat reduced by increasing the number of flagged peaks, but, surprisingly, this is not true for the LMWS (Figure 2). Apparently the peak flagging makes possible the retrieval of new correct answers to an extent which is proportionately large in comparison to the new wrong answers retrieved.

Maximum Percent Contamination

This restriction was imposed on PBM to speed the search and increase the precision by eliminating matches that require impurities in concentrations higher than the interpreter feels are possible. The tests on the spectra of pure compounds show that this only benefits search time; at higher precision values (Figure 1) the restrictions of 20% and 70% maximum percent contamination give nearly identical results (although a particular ΔK value is indicative of higher precision for the 20% than for the 70% runs). Note that the maximum recall value, which had been extended to >60% by the use of flagged peaks, is >90% with the more lenient restriction of 70% maximum percent contamination. Thus use of such a higher value is recommended even for

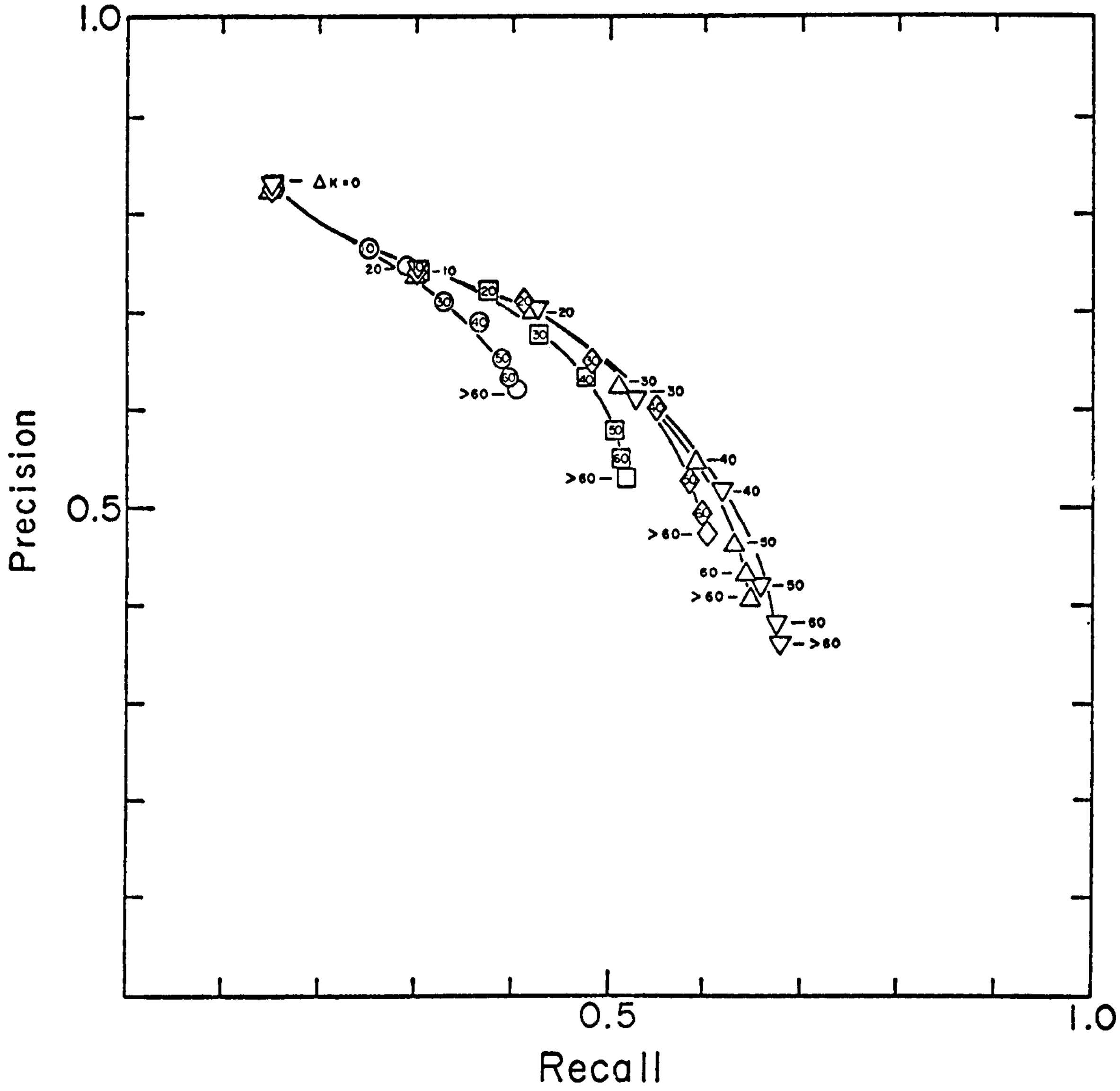

Figure 2. Effect of the number of flagged peaks on ΔK values for unknown LMWS spectra of pure compounds. Number of flagged peaks: O, 0; □, 1; ◇, 2; △, 3; ▽, 4.

unknown spectra of pure compounds; this helps for "incompatible data" (vide supra) of unknown peaks whose abundance is greater than predicted in the same way that flagged peaks help for those whose abundance is too low.

Class of Match

Adjusting the criteria for a structural match (Table 1) is equivalent to changing the relevancy decisions in a document retrieval system. Figures 3 and 4 clearly show that there is a very high probability that if a spectrum retrieved with a low ΔK value is not that of the compound identical to the unknown, it is actually that of a ring positional isomer, a homolog, or a compound whose structure differs only by the position of one carbon atom. In fact, for the higher K (or lower ΔK) values, most of the small proportion of remaining retrieved compounds even not in Class IV are of related structure, such as a dimethylhexadecanoic acid matched with octadecanoic acid. This behavior, which is also found for other retrieval systems (1, 3, 4, 6), shows that there are substantial cross-correlations of peak uniqueness values, as postulated in the explanation above of the precision achieved versus that predicted for a particular K value.

The relative effects of changing data classes on the results for the LMWS and the HMWS are significantly different: the largest proportion of the class I mismatches for the LMWS are ring positional isomers (Figure 3) whereas in the HMWS the homologs of the unknowns are the most significant (Figure 4). For example, at a ΔK of 30 for the LMWS two-thirds of the class I mismatches are ring isomers, half of the remaining mismatches are homologs, and about one-third of the remainder belong to class IV; for the HMWS at $\Delta K = 30$ well over half of the mismatches are homologs, while nearly half of the remainder are isomers which can be formed by moving only one carbon atom. This differing importance of the structural classes for the LMWS and the HMWS appears to be mainly an artifact of the makeup of the reference file. For example, the LMWS contains numerous dimethylnaphthalenes and dimethylindoles, with the positions of the two methyl groups occurring in various permutations on the rings; the spectra of these isomers are very similar. The HMWS contains a number of homologous long-chain aliphatic hydrocarbons and their derivatives such as primary alcohols and esters. Although the LMWS also contains many spectra of homologs, the peak abundances of these spectra are much more sensitive to the addition of a methylene group, the fragmentation patterns of methyl acetate and methyl propionate are easily distinguishable, while those of methyl heptadeconoate and methyl octadecanoate are nearly identical except in the molecular ion region. This can account also for the much larger effect of class IV data on the HMWS than on the LMWS, as a single "misplaced" carbon atom will tend to have a much smaller effect.

Molecular Ion Information (K+)

The molecular ion provides additional information which is especially valuable for distinguishing between homologs, as seen for the HMWS data in Figure 4. The increases in precision obtained by examining those class I matches retrieved with a K+ value are nearly commensurate with the increases obtained by using class III matching criteria with K values;

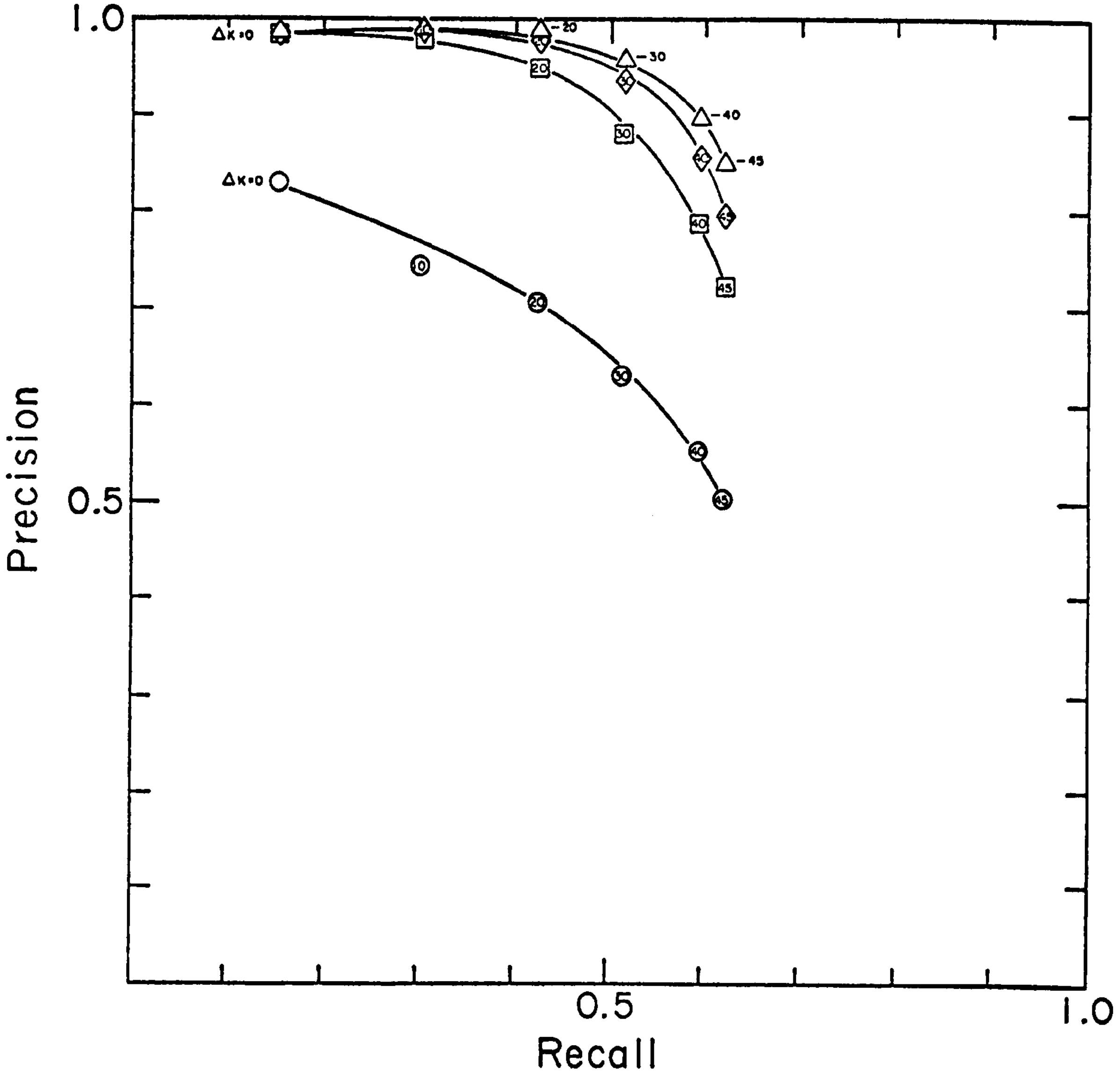

Figure 3. Effect of structural matching criteria on ΔK values for unknown LMWS spectra of pure compounds. Class of match: ○, I; □, II; ◇, III; △, IV.

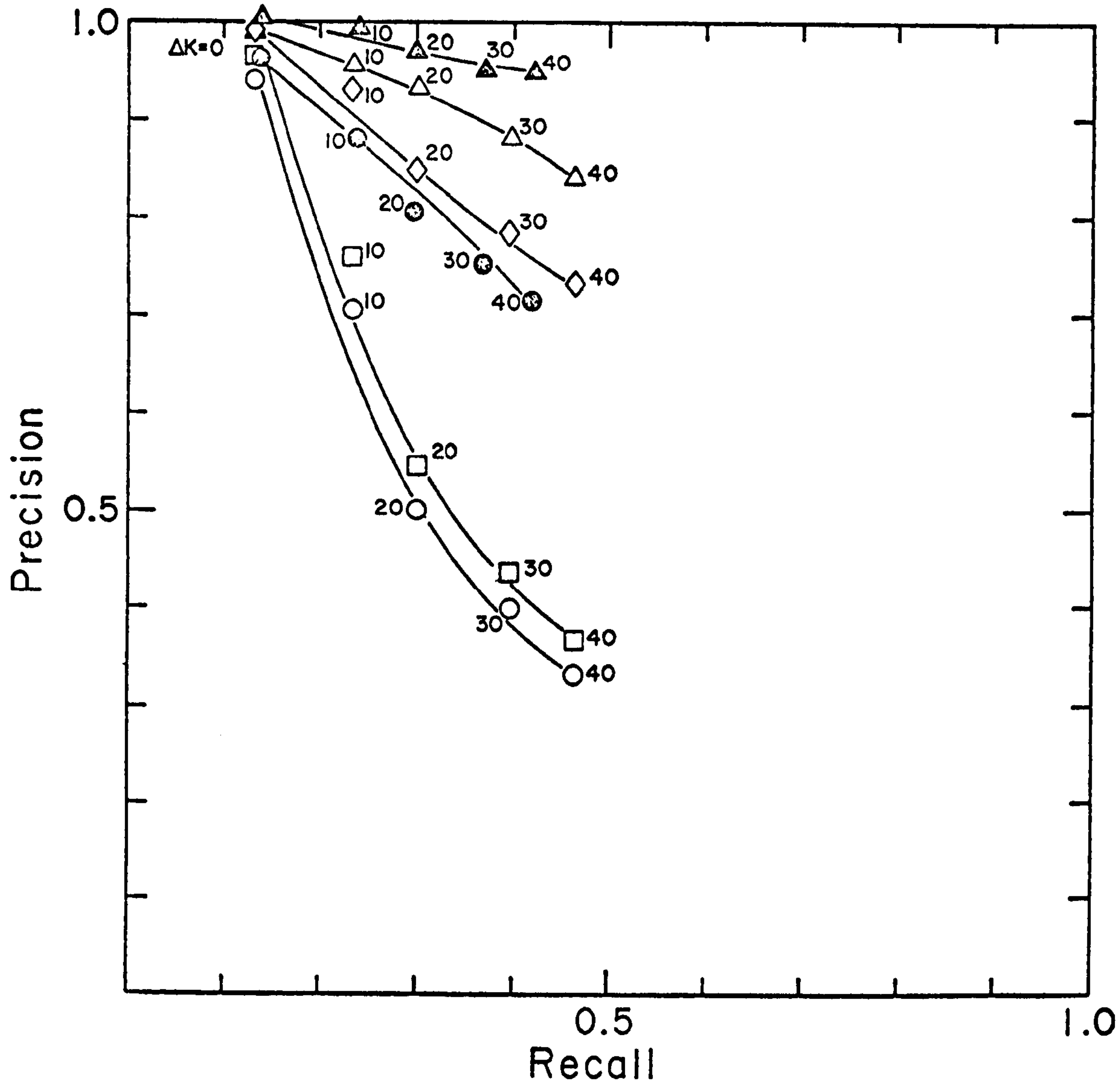

Figure 4. Effect of structural matching criteria and molecular ion information (K+) on ΔK values for unknown HMWS spectra of pure compounds. Class of match: ○, I; ⊛, I+; □, II; ◇, III; △, IV; ▲, IV+.

obviously the molecular ion should be uniquely effective in distinguishing between homologs. The same effect is significant in the consideration of class IV data as well. Thus for a high molecular weight unknown (Figure 4) a ΔK+ value ≤ 40 provides a 95% confidence of at least a class IV match, while for a low molecular weight unknown a ΔK+ value ≤ 30, which is obtained for nearly 50% of all possible matches, provides >99% confidence of a class IV match.[10] Note, however, that a high K+ value does not necessarily insure that the reference compound has the same molecular weight as the unknown. The molecular ion of a lower molecular weight compound can occur as an odd-electron fragment ion in the spectrum of a higher molecular weight unknown; matching the molecular ion is a necessary but not sufficient condition to prove identical molecular weights. The LMWS data[12] indicate that the K+ values are of little benefit in distinguishing between ring positional isomers (class II), as would be expected.

K Versus ΔK Values

The recall/precision performances using K and ΔK values show appreciable differences. For the HMWS (Figure 1) the precision achieved using K values is superior for recalls of 50 - 80%, while the opposite is true for the LMWS[12] for this recall range. Because the best ΔK value (zero) is the same for all reference spectra, at this value 12 - 15% of the possible matches are already retrieved; higher precision can be achieved for the LMWS[12] using K values ≥ 100 (for K ≥ 100, no mismatches and 4% recall were found for the spectra studied). These precision results at low recall levels are based on samples that are small statistically; for the HMWS (Figure 1) the <u>decrease</u> in precision at the highest K values is probably an artifact of the small data set.[2] Here the observed 50% precision is due to the fact that one of the two spectra retrieved at K ≥ 130 is a mismatch, the spectrum of hexachlorofulvene retrieved for hexachlorobenzene as an unknown (actually a match by class IV criteria); the close similarity of these two spectra has been pointed out.[13]

Precision Value as the Criterion of Match

The K and ΔK values found for a particular selected reference compound can thus give substantially different levels of confidence based on the recall/precision performance. Also (*vide supra*) the precision found for a particular value of K or ΔK is substantially dependent on the molecular weight, number of flagged peaks, maximum percent contamination, class of match, and inclusion of the molecular ion. Based on these recall/precision studies,[12] we are at present modifying PBM to convert the various types of K and ΔK values found for each reference spectrum to a "predicted precision" value which can be used in place of the K value for ranking the matches found, and which should provide a more direct measure of the confidence which the interpreter can place in the result with respect to each class of match.

Comparison of PBM with Other Systems

Of the variety of retrieval systems proposed which do not require human decision, that of Biemann and coworkers (the MIT system)[11] appears to be accepted as the one of best overall performance; the number of peaks employed (two for every 14 mass units of the spectrum) also requires substantially more computer time and storage than do other methods in general use.[1] Overall, the recall/precision performance for pure compounds of the PBM system is equivalent or superior to that of the MIT system (Figure 5) in the high precision (>50%) range, although inferior at low precisions; for mixtures (Figure 6) the PBM system is dramatically superior at all precision/recall levels. It should be noted that performance quality at high precision levels was a prime objective for PBM, as an unknown spectrum for which a match of high confidence could not be obtained should also be interpreted by a mass spectrometrist or a computer interpretive system such as STIRS.[10]

For the spectra of pure compounds in the LMWS, PBM gives clearly superior precision using ΔK criteria which recall 15% to 60% of the possible matches; precisions obtained using K values approach those of the MIT system for recalls <15%. Note that the MIT system uses ~30% more peaks for matching in the LMWS range and employs a forward search comparison. At low recall levels the MIT system is in turn clearly superior to PBM. This is in part due to the larger number of peaks and possibly the forward search mode of the MIT system; it is also in keeping with the tighter abundance criteria demanded by PBM which make spectra taken under substantially different experimental conditions irretrievable. It follows that relaxing the abundance criteria of PBM by increasing the window tolerances, an option available to the operator, should increase the recall at low precision values. Improved performance in this range should also be possible by "skewing" the unknown spectrum, either increasing or decreasing the observed peak abundances as a function of mass; this should compensate for instrumental mass discrimination or for changing sample concentration during spectral recording which has occurred for either the unknown or reference spectrum.

For the spectra of pure compounds in the HMWS (Figure 5) the MIT system performance again is substantially superior at low precision values. For precision values >50%, the PBM recall using ΔK values is closely equivalent to the MIT retrieval performance; using K values (Figure 1) the PBM recall performance is actually superior for 50% to 80% precision. However, the MIT system uses <u>twice</u> as many peaks for matching the HMWS range; because the PBM performance is degraded much more than that of the MIT system with increasing molecular weight, we have increased the number of peaks for the system presently used: for the molecular weight range beginning at 170 amu, 16 peaks; 180, 17; 195, 18; 215, 19; 240, 20; 270, 21; 305, 22; 350, 23; 420, 24; 500, 25; and ≥600, 26. Results using this modified PBM system with 35,828 reference spectra are shown in Table 3. For an unknown spectrum made by combining 90% methyl stearate and 10% methyl oleate, the 22 spectra retrieved with highest K values were either correct answers or closely related molecules; note that the correct compounds, but not homologs, have been retrieved with K+ values.

Relaxing the matching criteria (Table 2) affects the MIT system performance[12] for pure compounds in a manner that is closely similar to that found for the PBM performance (Figures 3 and 4); for example, using class IV criteria (Table 2) 27% of the possible LMWS matches can be recalled by the MIT system with 95% precision,[12] which compares to 53% recall for PBM (Figure 3). This supports the proposal that the differences observed for the LMWS

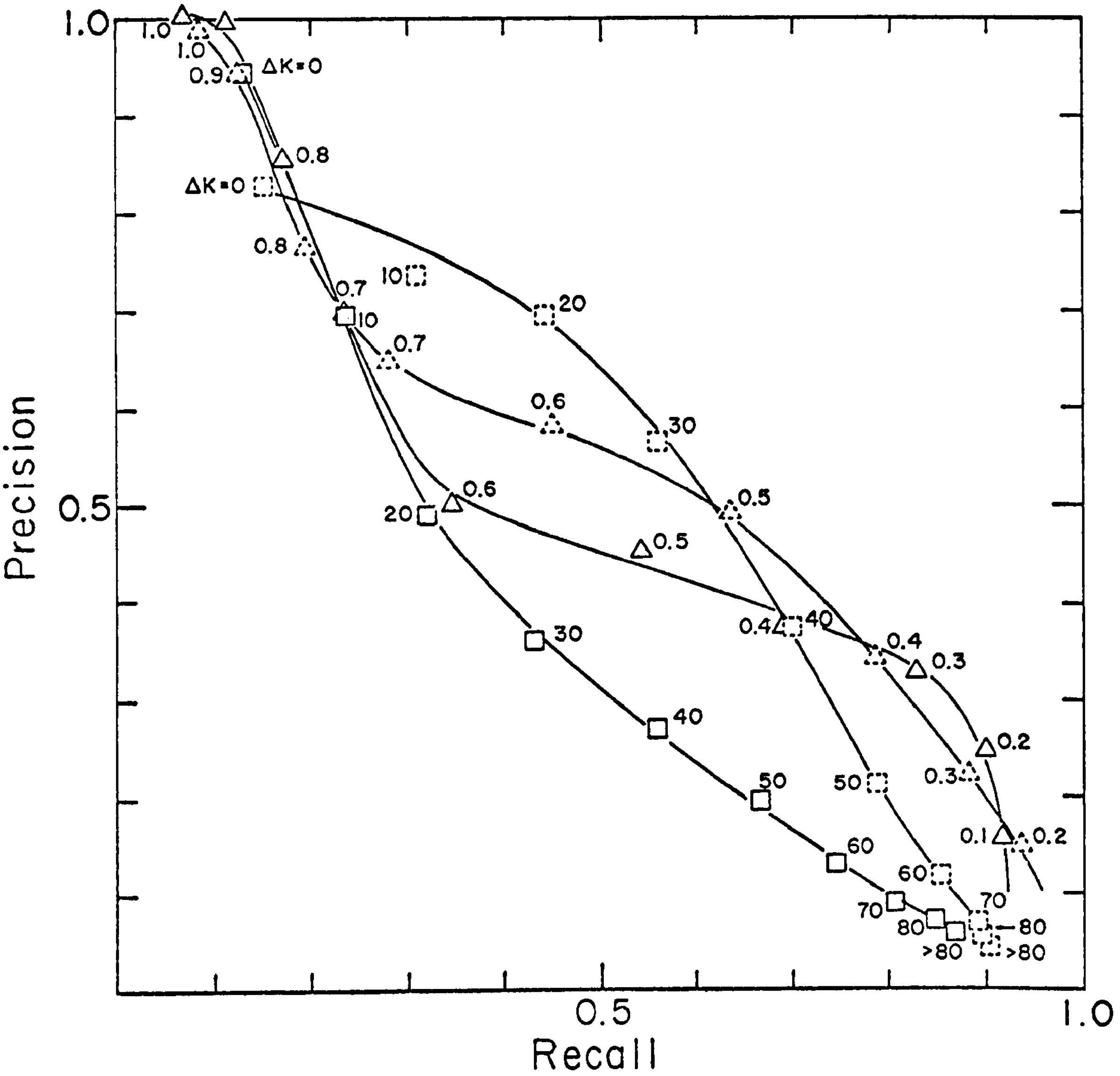

Figure 5. Comparative performance of retrieval systems for unknown spectra of pure compounds. LMWS: △, MIT; □, PBM. HMWS: △, MIT; □, PBM.

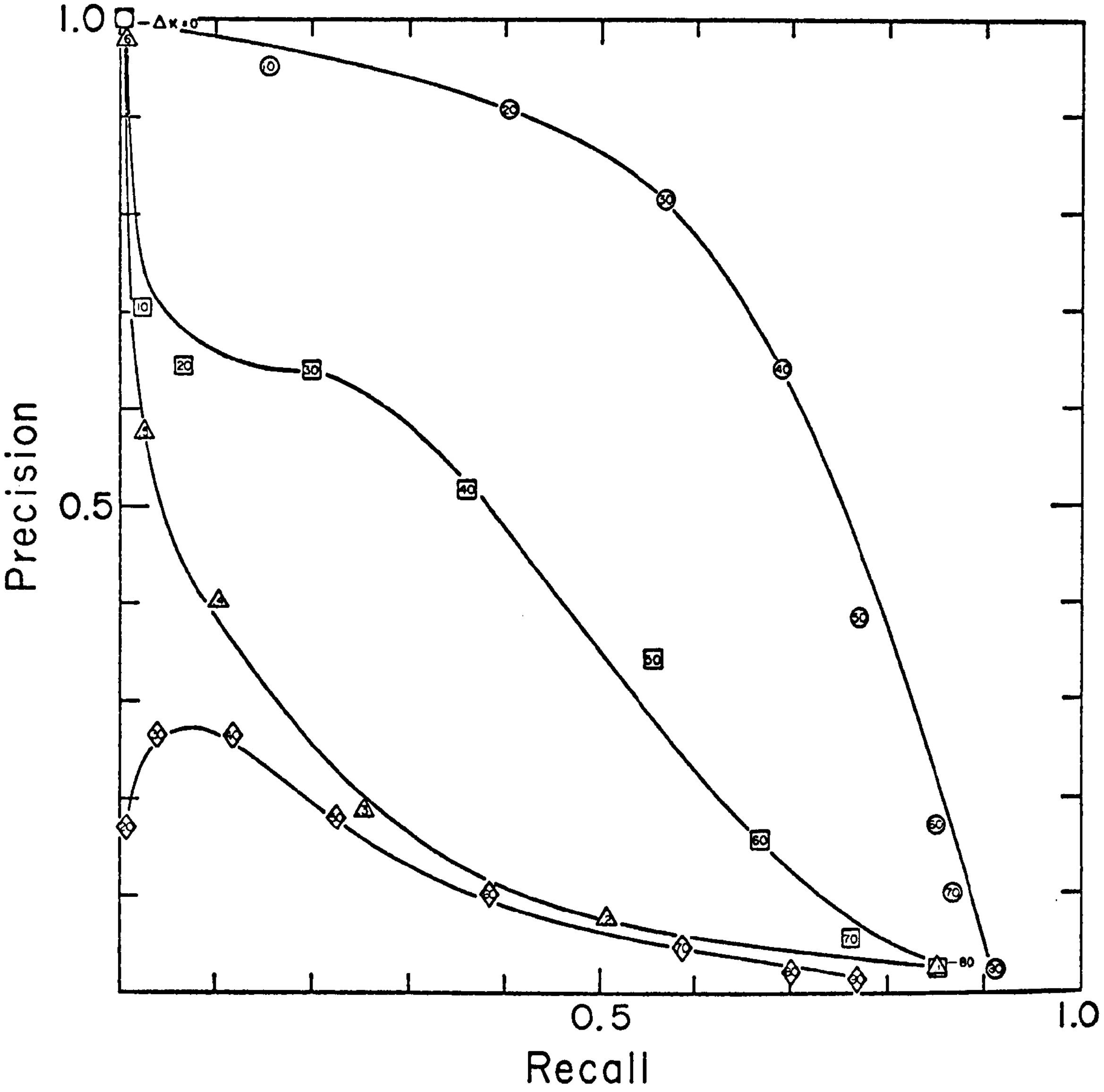

Figure 6. Comparative performance of retrieval systems for unknown LMWS spectra of mixtures. System and proportion of component present: △, MIT - 60%; ○, PBM - 60%; □, PBM - 30%; ◇, PBM - 10%.

TABLE 3. COMPOUNDS RETRIEVED BY PBM[a] FOR A MIXTURE OF 90% METHYL n-OCTADECANOATE + 10% METHYL cis-9-OCTADECENOATE

Compound	Confidence value		Percent contami- nation	Percent component
	K	ΔK		
Methyl n-octadecanoate,[b] $C_{19}H_{38}O_2$	134+, 95, 92+ 90**+, 78+	0, 7, 10, 12, 24	2, 27, 17, 22, 24	90, 76, 67, 91, 53
Methyl behenate, $C_{23}H_{46}O_2$	112*, 79, 61	0, 23, 41	27, 30, 66	63, 54, 23
Methyl 16-methylheptadecanoate, $C_{19}H_{38}O_2$	103+	0	23	65
Methyl arachidate, $C_{21}H_{42}O_2$	101, 77*, 73**	1, 25, 29	30, 27, 46	54, 56, 27
Methyl heneicosanoate, $C_{22}H_{44}O_2$	101**	1	27	87
Methyl nonadecanoate, $C_{20}H_{40}O_2$	95**, 65	7, 37	27, 50	78, 56
Methyl cis-9-octadecenoate[b]	85*+, 62**+	17, 40	84, 91	14, 10
Methyl 13,14-dideuteriooctadecanoate	81**	21	24	73
Methyl myristate, $C_{15}H_{30}O_2$	73**, 71**	29, 31	32, 34	100, 97
Methyl palmitate, $C_{17}H_{34}O_2$	70**	32	34	80
Methyl heptadecanoate, $C_{18}H_{36}O_2$	66**	36	36	61

[a] PBM specifications modified to include >15 peaks for reference compounds of molecular weight >170; 35,828 reference spectra searched.

[b] Correct answer.

and the HMWS in changing these criteria are largely artifacts of the reference file composition.

Application to Spectra of Unknown Mixtures

The precision achievable by PBM is dramatically superior for the spectra of unknown mixtures, with the differential improvement over the performance for pure compounds being attributable to the use of reverse searching.[3,4] Both the LMWS (Figure 6) and HMWS[12] are similar in showing recall/precision performance by PBM for components present in 30% concentration that is substantially above that for the MIT system with 60% components, which performance is actually approached by PBM using 10% components. For the 30% components, the MIT system retrieved only 10% and 7% of the total possible matches for the LMWS and HMWS sets, respectively, and <2% of the possible matches for the 10% components. Although many potential matches are apparently rejected by the base peak prefilter of the MIT system,[11] it is not expected that relaxing this criterion will be particularly helpful; for this system it is recommended[14] that "the mathematical resolution of component spectra of mixtures is the most satisfactory means of identifying minor components."

The PBM performance for 60% components appears to be superior to that shown earlier for the spectra of pure compounds (Figure 1). Although different data sets have been used, this is due mainly to the fact that the spectra used in making up the unknown mixture spectrum were not eliminated from the reference file (the same mixture spectra were of course used in the MIT system evaluation). Relaxing the matching criteria improves the precision for mixture spectral retrieval in the same fashion as observed for the spectra of pure compounds.[12]

Mixture Examples

Tables 4 and 5 show the compounds retrieved using the MIT and PBM systems for spectra of "unknown" LMWS and HMWS mixtures. The first spectrum was created by combining the spectra of 3-methoxyindazole, carbon tetrachloride, and tert-butyl-3-ketobutyrate in a 60:30:10 proportion. In the top 15 matches[12] (the top 10 are shown) selected by the MIT system, only the major component, 3-methoxyindazole, is retrieved, ranking third and seventh in the output list. The PBM results show that this 60% component is identified with high confidence ($\Delta K+$ values corresponding[12] to >95% precision). Although the confidence associated with the 30% and 10% components is much lower, molecular ion information and no flagged peaks were utilized in retrieving the tert-butyl-3-ketobutyrate spectrum, so that the confidence of that match is much greater than the confidence in any of the incorrect retrievals, for all of which the use of flagged peaks was necessary for matching.

Table 5 presents the results for a mixture of the herbicide Siduron, 1-(2-methylcyclohexyl)-3-phenylurea (60%), 1,2'-binaphthyl (30%), and the insecticide Sumthion, O,O-dimethyl-O-(4-nitro-*m*-tolyl)phosphorothioate (10%). All of the similarity indices obtained by the MIT system are extremely low, so that the low precision observed is not surprising. On the other hand,

TABLE 4. COMPOUNDS RETRIEVED BY THE MIT AND PBM SYSTEMS FOR A MIXTURE OF 60% 3-METHOXYINDAZOLE, 30% CARBON TETRACHLORIDE AND 10% tert-BUTYL-3-KETOBUTYRATE

System and compound	Similarity index or K value
MIT:	
1-methyl-3-indazolone	0.35, 0.34
3-methoxyindazole[a]	0.32, 0.27
2-methyl-3-indazolone	0.28, 0.28
p-allylanisole	0.27
1-methoxy-4-(1-propenyl)-benzene	0.27
2-methyl-3(2H)-benzofuranone	0.26
1-allyl-4-methoxybenzene	0.24
PBM:	
3-methoxyindazole[a] (33%, 33%)[b]	92+, 83+
carbon tetrachloride[a] (66%, 66%)[b]	55, 41
tert-butyl-3-ketobutyrate[a] (96%)[b]	42+
1-methyl-3-indazolone (48%)[b]	34*+, 25*+
chloropicrin (66%)[b]	29**
4-amino-1-methyl-1,2,3-benzotriazole (71%)[b]	26*+
3-phenyleicosane (83%)[b]	26**

[a] Correct answer.

[b] Value of "percent contamination" (%C) found by PBM; note that (1 - %C) is only an approximation of the actual concentration of the component present.

TABLE 5. COMPOUNDS RETRIEVED BY THE MIT AND PBM SYSTEMS FOR A MIXTURE OF 60% 1-(2-METHYLCYCLOHEXYL)-3-PHENYLUREA, 30% 1,2'-BINAPHTHYL, 10% O,O-DIMETHYL-O-(4-NITRO-m-TOLYL)PHOSPHOROTHIOATE

System and compound	Similarity index or K value
MIT:	
alpha-ionone	0.08
1-(2-methylcyclohexyl)-3-phenylurea[a]	0.07, 0.06
3-methoxy-4-hydroxymandelic acid	0.06
1-ethoxymethyl-4-methylenecyclohexane	0.06
N-phenyl-N'-methylurea	0.06
bornylene	0.06
cyclofenchene	0.06
tricyclene	0.06, 0.05
γ-terpinene	0.06, 0.05
bis(2-chloroethoxy)methane	0.05
PBM:	
1,2'-binaphthyl[a] (51%, 55%)[b]	89+, 59+
1,1-binaphthyl (51%, 69%)[b]	83*+, 39*+
O,O-dimethyl-O-(4-nitro-m-tolyl)-phosphorothioate[a] (84%, 88%, 93%)[b]	83+, 38+, 35*
1-(2-methylcyclohexyl)-3-phenylurea[a] (61%, 61%)[b]	73+, 57**+
2,2'-binaphthyl (70%, 69%)[b]	44*+, 43+
α-phenyldibenzofulvene (53%)[b]	41+
3,4-benzpyrene (89%, 89%)[b]	37**, 37**
γ-terpinene (65%)[b]	39**+

[a] PBM specifications modified to include >15 peaks for reference compounds of molecular weight >170; 35,828 reference spectra searched.

[b] Correct answer.

all three components actually present are retrieved by PBM, and the other compounds selected are structurally similar.

The PBM results thus confirm the advantages of both the reverse search strategy[3,4] and the weighting of mass and abundance values of peaks[3] for matching unknown mass spectra. Reducing the number of peaks necessary to achieve relatively high information precision also yields a significant reduction in search time requirements; it should be possible to do such a PBM search in real time for GC/MS. For example, matching against a reference file of 1,500 spectra during quadrupole MS data acquisition and reduction by a DEC PDP-8 computer (16K words core, 1.6M words disc storage) should require ~2 sec for an unknown mass spectrum.

PERFORMANCE OF THE STIRS SYSTEM IMPROVEMENTS

Further details in discussion are available in the Ph.D. thesis of Dr. H. E. Dayringer.[15] Results of the substructure identification system will be discussed first, and then this system will be used to evaluate the performance of STIRS with a variety of data class revisions.

Substructure Identification Capability

The recall abilities at the 98% confidence level exhibited by MF 1-7, 10, and 11 for 18 selected substructural groups and the averages for all groups are shown in Table 6. Tables 6 and 7 show the information precision and recall values for MF 11 for 179 substructures at the predicted 98% confidence level. Table 8 lists the 25 substructural groups of the 204 tested for which STIRS was unable to identify the group in a single compound (zero recall) at the 98% confidence level. A separate, but more complex, system for estimating confidence levels was also developed which incorporated the actual match factor values of the selected compounds. Extensive tests on 27 substructures showed results comparable, but not clearly superior, to those of the random drawing model.

Information Precision

The overall STIRS results for 179 individual substructures using MF 11 show an average information precision of 98.1%, surprisingly close to the predicted value of 98%. The values vary from 100% down to 89% (xylene), with only three showing values <93%, and 23 <96%. The fluctuation in these values appears to be primarily statistical, giving strong support to the basic validity of the random drawing model to predict the confidence level of the STIRS substructural assignment. (In a number of cases, incorrect assignments were found to be due to errors in the spectra used as unknowns). The information precision values found for the 98% confidence predictions of the other data classes (MF 1 - MF 10) varied over a much wider range.

The MF 11 results at the 99.5, 99.8, and 99.9% levels show higher information precision (and lower recall) values, but the change was not as great as expected; even at the 99.9% confidence level the average precision only increased to 99.0%. Thus in this study we will not attempt to distinguish between information precision levels above 99%. Note that in the previously

TABLE 6. RECALL ABILITY (%) OF STIRS AT THE 98% CONFIDENCE LEVEL FOR NINE MATCH FACTORS

WLN[a]	Substructure[b]	% in file	MF1[c] Ion series m/e: 27-99	MF2 -Characteristic 27-89	MF3 ions- 90-149	MF4 ≥150	MF5 -Neutral 0-64	MF6 losses- ≥65	MF7[d] II°	MF10 2 peaks/ 14 amu	MF11 Overall, ΣMF1-6	Information precision, MF11
≥3	alkyl ≥C_3	20.3	24	43	28	11	36	19	10	33	56	90.2
V	carbonyl	46.2	1	13	11	6	27	12	5	18	31	100.0
Q	hydroxyl	20.3	15	25	15	4	36	5	10	12	42	93.4
VO	ester, anhydride	19.4	5	21	16	11	49	18	9	32	55	96.4
R	phenyl	28.4	30	54	40	16	27	25	7	44	75	93.7
Z	amino, amido	4.3	6	31	13	0	19	6	13	25	44	98.1
M	imino, imido	11.9	12	7	14	9	14	2	2	35	19	91.0
N	trisubst. N	28.8	22	27	18	11	24	11	14	22	30	95.3
S	sulfur	10.4	16	33	14	4	22	10	22	27	35	98.2
O1	$-CH_2O-$, CH_3O	19.0	3	25	18	12	35	15	6	24	49	94.8
L..J	carbocyclic ring	24.1	30	46	43	14	20	10	8	46	50	94.7
F	fluorine	5.0	0	39	28	6	50	33	28	22	61	99.0
G	chlorine	8.2	17	29	17	9	71	17	37	29	74	99.6
E	bromine	3.4	6	11	17	17	39	56	28	22	67	97.9
I	iodine	1.1	10	23	30	33	20	30	10	47	47	93.4
-SI-1&1&1	trimethylsilyl	4.5	43	90	63	50	67	47	7	63	97	97.7
L6TJ	cyclohexane	3.1	47	73	63	30	47	27	10	70	80	95.6
QVR	benzoic acid	0.8	7	43	27	13	43	7	20	53	73	98.5
Number of functionalities giving non-zero recall values			112	167	158	136	159	127	125	169	179	98.1[e]
Average recall for these			15	32	28	22	26	18	14	40	49	

[a] In many cases additional WLN permutations were used as descriptors of a particular substructure; most of these are listed in reference 17; a complete list is available from the authors.

[b] Approximate description of the group defined by the WLN symbol; see text for possible ambiguities.

[c] Complete descriptions of these mass spectral data classes are contained in reference 2.

[d] Match factors 8 and 9 did not satisfactorily identify any of the functional groups.

[e] Average precision of MF11 results for 179 substructures giving non-zero recall.

TABLE 7. RECALL AND PRECISION OF STIRS MF11 PREDICTIONS AT THE 98% CONFIDENCE LEVEL FOR 161 OTHER SUBSTRUCTURES

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
1	$-CH_2-$, CH_3	8886	98	25
2	$-CH_2CH_2-$, C_2H_5	3793	93	19
O2	$-CH_2CH_2O-$, C_2H_5O	975	94	42
O	linking oxygen	7709	99	39
U	double bond	4495	95	38
K	quaternary amine	138	96	50
Y	single branch	4985	94	28
≥3O	alkoxy, $\geq C_3$	494	94	50
P	phosphorus	424	98	70
T..J	heterocyclic ring	6796	97	31
VS	thioester	68	100	43
UU	triple bond	216	96	29
T56 BN DOJ	benzoxazole	20	100	55
X	double branch	2039	93	44
MV1	acetamido	202	97	60
-SI-	silicon	950	98	90
U1V1	acetonylidene	53	98	37
OVR	benzoate	387	99	53
1V	acetyl, $-CH_2CO-$	2159	95	47
VQ	carboxy	597	98	43
OV1	acetate	1100	96	47
L66BG AB-C 1B ITJ	adamantyl	15	100	67
GV	acid chloride	57	99	50
VYZ	alanyl	24	99	25
V1U1	acrylyl	167	96	50
1U2	allyl	312	99	23
V4V	adipyl	21	100	19
ZR	aniline	115	97	30
VH	aldehyde	424	97	37

Table 7. Continued

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
1OR	methoxyphenyl	829	97	43
VZ	primary amide	182	97	57
L C666J	anthracene	14	100	14
MR	N-subst. aniline	397	96	43
NNN	azide	95	99	53
1OR XV	anisoyl	105	98	50
T3MTJ	aziridine	46	99	73
L C666 BV IVJ	anthraquinone	16	99	19
NO&UN	azoxy	7	100	57
UNNU	azino	44	97	30
VHR	benzaldehyde	92	99	83
NUN	azo	83	97	77
L C6 B666J	benzphenanthrene-3,4-	9	100	89
L57J	azulene	6	100	50
T56 BM DNJ	benzimidazole	45	99	53
ZVR	benzamide	21	98	48
T56 BOJ	benzofuran	21	100	33
Q1R	benzyl alcohol	106	95	20
T56 BSJ	benzothiophene	35	99	70
U1R	benzylidene	377	94	50
L66 A BTJ	bicyclo(2.2.2)octane	40	98	70
RYR&U	benzohydrylene	35	98	53
L4TJ	cyclobutane	25	100	36
T56 BN DSJ	benzothiazole	36	100	37
L55 ATJ A A	bornane	19	99	47
RR	biphenyl	19	100	79
T5OVTJ	gamma-lactone	36	100	67
O4	butoxy	184	98	53
V1U1R	cinnamoyl	94	98	77
L7TJ	cycloheptane	11	99	45
T66 BOVJ	coumarin	45	100	27

Table 7. Continued

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
T56 BOT&J	coumaran	16	99	19
CN	cyano	307	98	37
V1U2	crotonyl	29	100	17
L6U CUTJ	cyclohexadiene-1,3-	22	100	27
L5 AHJ	cyclopentadiene	9	100	33
L6VTJ	cyclohexanone	48	99	37
L5VTJ	cyclopentanone	25	100	36
L5TJ	cyclopentane	146	97	60
NW	nitro	508	99	77
L3TJ	cyclopropane	89	100	57
N1&1	dimethylamino	325	97	63
L66TJ	decalin	37	98	73
SW	sulphonyl	254	97	53
UNMR BNW DNW	2,4-dinitrophenylhydrazone	32	100	87
L B656 HHJ	fluorene	18	100	39
SS	disulphide	80	100	47
T5OJ	furan	110	99	43
MVH	formamido	14	100	14
Q1V	glycolyl	26	99	23
T5NNOVJ	sydnone	36	100	67
MZ	hydrazino	63	99	23
T5MVMV EHJ	hydantoin	11	97	18
QM	hydroxyamino	54	99	30
V2R	phenylpropionyl	22	100	50
T56 BMNJ	indazole	18	100	50
L56T&J	indan	53	98	83
T56 BMT&J	indoline	12	100	63
T56 BMJ	indole	134	98	63
QNU	isonitroso	43	100	23
RMNU	phenylhydrazone	53	100	73
T5NOJ	isoxazole	64	99	43

Table 7. Continued

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
SCN	thiocyanate	18	100	11
SH	thiol	125	99	57
V1V	malonyl	75	98	23
1S	methylsulfide	306	97	33
O1O	methylenedioxy	14	100	29
L E5 B666..J	steroid skeleton	782	97	87
T6M DOTJ	morpholine	38	100	27
NO	nitroso	508	99	70
L66J	naphthalene	167	97	77
W	oxalyl	97	96	20
L55 ATJ	norbornane	65	99	70
T5N COJ	oxazole	35	99	83
L E5 B666 LUTJ	steroid skeleton, 6,7-dehydro-	99	98	70
OO	peroxy	346	96	60
QNU	oxime	43	100	23
U1VR	phenacylidene	29	98	52
1VR	phenacyl, acetylphenyl	149	93	43
MVMR	phenylureido	24	98	63
L B666J	phenanthrene	34	99	70
QR	phenol	957	97	70
T C666 BN INJ	phenazine	37	100	47
OR	phenoxy	1464	94	40
T C666 BM ISJ	phenothiazine	65	99	60
NUNR	phenylazo	72	97	90
QV1R	phenylacetic acid	15	99	80
T6M DMTJ	piperazine	30	100	60
T56 BVMVJ	phthalimide	27	99	37
3U	propylidene	764	97	30
V2	propionyl	393	99	40
T6N DNJ	pyrazine	50	100	67
T66 BN DN GN JNJ	pteridine	26	100	73

Table 7. Continued

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
T6NJ	pyridine	278	99	63
L666 B6 2AB PTJ	pyrene	6	100	50
T6N CNJ	pyrimidine	81	100	63
T6NJ AO	pyridine-N-oxide	9	100	78
T66 BNJ	quinoline	111	99	77
T5MTJ	pyrrolidine	79	96	20
L6V DVJ	quinone	32	100	37
T66 BN DNJ	quinazoline	38	99	80
QR BV	salicyloyl	62	98	67
T66 BN ENJ	quinoxaline	30	100	53
1U1R	styrene	176	96	70
UNMVZ	semicarbazono	41	100	83
V2V	succinyl	16	96	6
T5VNVTJ	succinimide	12	100	8
T5OTJ	tetrahydrofuran	124	97	60
MSW	sulfonamido	38	99	50
T5N CSJ	thiazole	25	100	20
T6OTJ	tetrahydropyran	745	98	77
US	thiocarbonyl	203	97	60
SUYZ	thiocarbamyl	16	100	25
T5SJ	thiophene	158	99	77
NCS	thiocyanate	18	100	11
R1V	phenylacetyl	71	97	70
MR X1	toludino	61	98	37
T6N CN ENJ	triazine	59	99	80
1R X	xylene	440	89	30
ZVZ	urea	306	98	47
FXFF	trifluoromethyl	343	99	63
1U1	vinyl	557	94	23
4V	valeryl	126	96	30
RV	benzoyl	1039	94	53

Table 7. Continued

WLN[a]	Substructure[b]	No./18806 in file	Information precision	Recall
T C666 BO IVJ	xanthone	9	100	56
1V1V	acetoacetyl	36	99	27
T56 ANJ	indolizine	24	100	63
T7MVTJ	caprolactam	8	99	50
MNW	nitramino	7	99	29
T5MN DNJ	triazole, 1H-1,2,4	26	100	23

[a] In many cases additional WLN permutations were used as descriptors of a particular substructure; most of these are listed in reference 17; a complete list is available from the authors.

[b] Approximate description of the group defined by the WLN symbol; see test for possible ambiguities.

TABLE 8. SUBSTRUCTURES OF ZERO RECALL BY MF11

WLN[a]	Substructure[b]	No./18806 in file
ZR BV	anthraniloyl	12
MVR	benzamido	40
SHR	benzenethiol	12
UU1R	benzylidyne	13
L35TJ	bicyclo(3.1.0)hexane	4
T B656 HMJ	carbazole	5
T B656 EN HMJ	carboline, beta	5
L7 AHJ	cycloheptatriene	8
U2U1R	cinnamylidene	4
T B656 HOJ	dibenzofuran	5
T B656 HSJ	dibenzothiophene	4
V3V	glutaryl	7
VHV	glyoxylyl	4
V1MVR	hippuryl	5
L56 BHJ	indene	4
OCN	isocyanato	9
T66 CNJ	isoquinoline	9
T5VOVJ	maleic anhydride	4
L46 ATJ	norpinane	4
T40TJ	oxetane	9
T66 CNNJ	phthalazine	8
1UU1V	propiolyl	6
QR B1U	salicylidene	12
T E5 B666..J	steroid, heterocyclic	19
SWQ	sulfonic acid	5

[a] In many cases additional WLN permutations were used as descriptors of a particular substructure; most of these are listed in reference 17; a complete list is available from the authors.

[b] Approximate description of the group defined by the WLN symbol; see test for possible ambiguities.

cited example of 373 spectra of which 70 contained the substructure, and STIRS identified its presence in 35 of the 70 (correctly) and 3 of the remaining 303 (incorrectly), a decrease of only one in the incorrect identifications increases the precision value from 98.0% to 98.7%. Further, and more importantly, a sampling of the incorrect identifications by MF 11 shows that most actually contain substructures that are closely similar, at least on the basis of their mass spectral behavior, to that identified. For example, one-third of the compounds incorrectly identified as containing a phenyl actually had a benzo group, and nearly half a pyridine ring; in many cases these alternative identifications were indicated by STIRS, either directly as that substructure or as a functionality of common occurrence in the compounds selected with highest MF values. For the VO classification, -CO-O-, for a large proportion of the incorrect identifications the WLN of the compound contained VQ, -CO-OH. Although this results from the necessarily arbitrary nature of some classifications of the substructures, in the case of a true unknown such "incorrect" identifications could be helpful, or at least not be seriously misleading.

The gratifying agreement between the average MF 11 information precision value of 98.1% and the predicted confidence value of 98% also is a strong indication that the validity of the STIRS results are not significantly dependent on the data base, unknowns, or substructures. The substructures were not chosen on the basis of their mass spectral behavior (although this should be beneficial, vide infra), so that a similar STIRS performance would be expected for new compounds even of unusual mass spectral behavior examined as unknowns or added as reference spectra to the file. However, because STIRS can only identify structural moieties which are already in the reference file, it should be remembered for a total unknown that the substructures indicated could thus be only those closely related in their mass spectral behavior. For example, if an unknown had a steroid structure except that the "D" ring was six-membered, and there were no reference spectra of this type in the file, STIRS might well indicate the steroid substructure at a high confidence level (although the actual MF values could be lower than expected). However, if this compound type was also well-represented in the reference file, but was not in the substructure list, these reference compounds should then appear instead in the top 15 selected, and thus reduce the confidence of the steroid substructure prediction.

In summary, the information precision values found here indicate that the STIRS identifications from MF 11 are reliable with the predicted confidence at least to the 99% level, i.e., if STIRS predicts the presence of a pyridine ring in an unknown with 99% confidence, in only one case of 100 should this turn out not to be true.

Recall

To reemphasize, although it is important that STIRS be correct in a high proportion of the cases that it makes a substructure prediction (a high "information precision"), it is also important that it makes such a high confidence prediction of the particular substructure in a substantial proportion of the unknowns in which the substructure is actually present (a high "recall"). Considering the wide range of structural types, the average recall by MF 11 of 49% for 179 substructures appears to be a promising performance, and values such as 97% for trimethylsilyl (identifying the group in 29 of the 30 TMS compounds examined), 89% for benzphenanthrenes, 87% for steroids, and 87% for dinitrophenylhydrazones are quite impressive. For 12% of the substructures examined (Table 8), STIRS gave zero recall, but for all but one of these (MVR, benzamido) there were only thirteen or less compounds which contained that substructure in the file of 18,806 compounds.

The utility of the system depends, fundamentally, on the amount of reliable information which it can supply on the average unknown molecule. Multiplying the recall for each substructure by its proportion in the data base, and summing these values, gives a figure of 2.55; this means for an average unknown spectrum STIRS should be able to identify two or three substructures by MF 11 with high confidence. Of course this number will increase if the list of substructures is made more comprehensive or if the STIRS performance is improved to increase the recall values.

In almost every case the overall match factor, MF 11, showed a higher recall value than any of the other data classes. This is not surprising, as MF 11 is a weighted average of MF 1 - 6 (1 x MF1, 1 x MF2, 2 x MF3, 2 x MF4, 4 x MF5, 2 x MF6), and in most cases MF 7 - 9 were of marginal utility. In general the variation in recall values with data class (MF 1 - 11) and substructure corresponds to know spectra-structure correlations, as noted earlier for the qualitative information from STIRS.[10] Those substructures giving characteristic fragmentation behavior generally show higher recall; the high MF 2 recall for phenyl is consistent with the characteristic "aromatic ion series" in the m/e 37 - 79 mass range, while the high MF 5 recall for chlorine corresponds to the high tendency for the loss of this electronegative species as neutral Cl or HCl (note that none of the data classes should be sensitive to the characteristic isotopic abundances of chlorine). Trimethylsilyl derivatives exhibit abundant ions at unusual masses such as 73, 75, 89, and 147, making the identification of this substructure possible in a high proportion of spectra.

Despite the substantially superior performance of MF 11 in general, for a particular unknown spectrum it is possible that one or more substructures can be identified at a higher confidence level by another data class (MF 1 - MF 10). It is important, however, that predictions be made only for those substructure/match factor combinations of high information precision and recall. If each of the 204 substructures were to be predicted by each of eleven data classes, even if there were only a 1% chance that a particular match factor indicates a particular substructure because it is present by chance in the required number of the top 15 compounds, this would mean that approximately 22 substructures (1% x 11 x 204) would be indicated incorrectly for the average unknown. Thus in practice we use the MF 11 results as the primary structural information. Although it is helpful to examine the MF 1 - MF 10 results for other structural clues, the only substructure-MF combinations used are restricted by the computer to those for which the statistical study showed information precision values ≥94%, and recall values >30% of the value for MF 11.

Application to Unknown Spectra

The applicability and limitations of this technique are illustrated by STIRS substructure predictions (≥98% confidence level) for some "unknown" mass spectra. As discussed above, primary reliance is placed on the MF 11 results, but those from MF 1 - MF 10 are reported to show how these can supply confirmatory evidence and indicate additional possible substructures.

Results with simple molecules are generally excellent. From the spectrum of 5-tridecanone (WLN, 8V4) STIRS identified an alkyl group of three or more carbons (≥3) by MF 3 and MF 11, with no incorrect identifications. For methyl trichloroacetate (WLN, GXGGVO1) the substructures methoxy (O1, by MF 11), tetrasubstituted carbon (X, by MF 2, 3, 10), chlorine (G, by MF 5, 7, 10, 11), and "GV" (by MF 5, 11) were identified. Despite the fact that "GV" can indicate an acid chloride, ClCO-, this as well as all of the other STIRS identifications are actually correct; G is a terminator in WLN, and for all of the compounds selected by STIRS the chlorine was substituted on a carbon attached to the carbonyl group, not on the carbonyl itself. This emphasizes the importance of examining in addition the actual structures of the selected compounds, as well as careful definition of the WLN subcodes used (this subcode has now been narrowed to restrict this group to only acid chlorides by requiring that there be no structure symbol immediately preceding the GV).

From the spectrum of ethyl 2-isopropyl-3-oxobutyrate (WLN, 2OVYVI&Y) STIRS correctly identified six substructures from the MF 11 results with >99% confidence and gave no incorrect MF 11 assignments. Although some of the simpler ones (ethyl, linking oxygen) are redundant, the identification of groups such as VO (ester), O2 (ethoxy), Y (single branch), and 1V (acetyl) would be useful in elucidating the structure of this as an unknown. Of the secondary predictions by other MF data classes, two substructures are correct, and 1V1, V1V, and 1V1V (CH_3COCH_2CO-) predicted by MF2 are similar to the correct substructure CH_3COCH_2CO-)-. Three of these secondary predictions are misleading; carbocyclic (L..J, by MF 3), sydnone (T5NNOVJ, by MF 6), and naphthalene (L66J, by MF 3). However, if C_2H_5O-CO-C- and C_2H_5O-CO-CH($COCH_3$)- had been included in the substructure list, the STIRS results would have been much more dramatic; 11 of the top 15 compounds selected by MF 11 were ethyl esters, while <u>five</u> were ethyl esters of 2-alkyl-3-ketobutyrates.

From the spectrum of γ-lactone functionality (T5OVTJ, by MF 2, 11) as well as the alkyl chain (≥3, by MF 3, 11), carbonyl (V, by MF 10, 11), and ester (VO, by MF 11). The related group $-CH_2CH_2CO-$ (V2, by MF 10, 11) was the only other substructure found. The spectrum of δ-laurolactone also produced STIRS predictions of T5OVTJ, ≥3, V, and VO; the δ-lactone substructure is not included in the substructure list. The top 15 compounds selected by MF 11 actually contained two δ-lactones as well as the four γ-lactones on which the substructure prediction was based, while the STIRS results for the γ-lactone showed one δ-lactone as well as the four γ-lactones selected for MF 11. Thus STIRS does not differentiate well between the two substructures, making it preferable to combine these into a single substructure indicating either of these functionalities.

Utility of STIRS Substructure Identification

It should be emphasized that this system for substructure identification, as STIRS itself, is intended as an aid to, not as a replacement for, the human interpreter. The capabilities of STIRS are subject to the basic limitations of mass spectrometry. STIRS can give no more information from a particular unknown mass spectrum than a human interpreter having unlimited time, insight, and experience, and for a molecule of moderate complexity neither the interpreter nor STIRS can make a substructure prediction with absolute certainty (100% information precision). Thus it did not seem particularly useful, or feasible, to conduct a statistically valid comparison of the abilities of human interpreters versus STIRS. In our applications of the method to date there have been many instances of STIRS substructure identifications by MF 11 at a high confidence level that were missed by manual interpretation. However, if the human interpreter did not recognize, for example, a sydnone substructure, it might be argued that someone with sufficient experience <u>in the spectra of such compounds</u> would not have missed this. Even for cases in which this is true, the increase in speed and confidence of the interpretation process provided by the STIRS substructure information would appear to justify the relatively small effort required to obtain this information.

Further comment may also be appropriate for the information precision and recall values observed here. The interpreter can consider substructures indicated by STIRS at the 80% confidence level, but must keep in mind that postulation will be incorrect in one of five cases. Of course it is much better to consider first substructures predicted by STIRS with higher precision values, although there will be fewer of these (lower recall). We contend that the average recall of 49% at the 98% confidence level for this variety of substructures indicates that this method has substantial utility; however, this is not based on a comparison with the performance of a trained mass spectrometrist, but rather on substantial experience in the degree to which STIRS substructure identifications can help any, including a highly experienced, mass spectrometrist. For example, an interpreter might be able to surpass STIRS in the identification of chlorine and bromine in unknown spectra, as none of the STIRS data classes utilize directly the isotopic abundance information characteristic of these elements. However, in cases in which the isotopic information is ambiguous because of interference or sensitivity problems, the interpreter might still be helped by a STIRS identification of these elements.

Finally , the quantitative evaluation provided by the information precision/recall values should be valuable for measuring the improvement achieved by further modifications to STIRS, and for performance comparisons with other systems (including those utilizing human as well as computer interpretation). For example, a very recent study by Kent and Gäumann for substructure identification from mass spectra using learning machine methods[16] gives performance data for twenty simple functionalities; useful identifications were not possible for compounds containing more than one type of substructure. Complete elucidation of structure, the ultimate goal in interpreting an unknown mass spectrum, utilizies a variety of spectral information such as molecular weight and ion elemental compositions from isotopic abundances in addition to substructure information. For complex molecules for which the complete structure cannot be determined at a high confidence level, it would appear to be helpful to know at least what parts of the structural assignment can be given with high confidence.

System Improvements

STIRS using MF 11 can identify carbonyl (V) with 31% recall, a performance which is encouraging in view of the scarcity of specific mass spectral peaks characteristic of this group. However, the recall is substantially higher for terminal carbonyl-containing functionalities such as acetyl (1V, 47%) and benzoyl (RV, 53%). Thus as might be expected, functionalities which are known to have a strong directing effect on mass spectral fragmentations, such as saturated nitrogen, can exhibit a low recall if defined in too general terms; the performance for the secondary amine (M) and benzamido (MVR, Table 8) substructures should be improved by substituting a number of more specific classifications such as $-CH_2NHCH_3$ (-1M1) and $-CH_2NHCOC_6H_5$ (-1MVR). Because the search for the present 204 substructures only increases the time for a STIRS run by a few percent, an obvious possible improvement is to subclassify each of the more general substructures of the 204 into a number of new substructures expected to give the most characteristic specific effects on the spectrum. In particular cases such as for the isopropyl group, substructure definition for the reference spectra by WLN is difficult, and definition from a connection table (or manually) will be preferable. The connection table is a computer representation of the atom connectivities of the molecule, and can be generated from the WLN.

With a few exceptions, the substructures tested here are relatively simple. An obvious extension of the method would be to include more complex substructures, especially those which are well represented in the reference file. For example, although a substructure for the steroid skeleton (WLN, L E5 B666..J) is included, the subset of estrogens (steroids with an aromatic "A" ring) is not. When the spectrum of 17-vinylestradiol 3-methyl ether was examined by STIRS, all of the top ten reference spectra selected by MF 11 were of estrogens;[17] this would predict that this relatively complex substructure should show a high recall value. More complex substructures might also be identified by having the computer search for combinations of substructures in the selected compounds.

Improvements in the information precision/recall performance should also be possible by modifications to STIRS, such as new match factors incorporating different data classes, and new "overall" match factors involving different combinations of the individual match factors. We would then plan to use for the prediction of a particular substructure only the few match factors which show the highest precision/recall performance.

STIRS Data Class Improvements

The recall and precision values found for STIRS predictions of substructure identity at the 98% confidence level are shown in Table 9. The statistical validity of these results is supported by the close correspondence to those of the original data classes which employ similar criteria. For example,

TABLE 9. PERFORMANCE OF DATA CLASSES FOR 98% CONFIDENCE LEVEL PREDICTIONS

Data class	Recall, %	Precision, %	Number of substructures with non-zero recall
2[a]	31.6	96.1	167
3[a]	27.8	95.7	158
4[a]	21.5	93.0	136
2A	32.7	95.5	174
2A'	33.4	97.3	171
3A	34.6	96.1	173
3B	31.2	96.0	169
4A	26.4	93.7	163
4B	18.6	91.7	139
4C	16.0	87.5	99
11[a]	49.1	98.1	179
11.1	47.4	97.9	183
11.2	47.3	98.1	182

[a] Data from reference 1.

the MF2A values are derived from data very similar to that used for the original MF 2 values, and these give comparable results in precision, recall, and number of identifiable substructures. The relatively smooth change in average recall found through each series of data classes, such as 2A-4C, indicates that there are no serious inconsistencies in the limits or extent of the mass ranges used. Because the contribution of a particular data class to the overall STIRS performance is also dependent on the number of substructures for which it can give the best recall, such data have also been determined (Table 10). These data give a very different impression of the usefulness of the data classes employing high mass peaks; although MF4C values give an average recall of 16% on 99 substructures, only a very small percentage of these show a recall ability that is superior to that of other data classes. Thus the additional structural information supplied by MF4C values does not appear to justify the computer time and storage required for its use.

Number of Ions

Changing the number of ions used in matching has a relatively small effect on the performance of a STIRS data class; using more peaks than the number previously recommended (Table 1) slightly increases the recall for substructures without substantially affecting the precision values. The mass ranges of data classes 3 and 3B correspond closely; thus the increase in average recall from 27.8% to 31.2% for predictions by MF3 versus MF3B results mainly from increasing the number of peaks employed from five to eight. In a study of 20 commonly occurring substructures it was found that in using the six (MF 2), eight (MF2A), and ten (MF2a) largest even-mass and odd-mass ions in the low mass range (Table 5) the average recall went from 26.0 to 28.2 to 29.7%, respectively, while the precision stayed the same (89-90%) versus the predicted 98% confidence level. However, for half of the substructures the use of ten peaks (MF2a) instead of eight (MF2A), gave the same or a lower recall value, indicating that in some cases the use of more information for matching can even "confuse" the substructural identification. Because additional computer time and space are required for each additional peak matched, it was decided for further testing to increase only to eight the number of ions used in the lower mass ranges (MF2A, 3A, 3B).

Varying the number of ions used in the higher mass ranges give similar trends in the recall values, but indicates that fewer peaks are necessary per mass increment. For the mass range 117-200 (MF4A) the average recall increases from 21.7% to 25.2% to 26.4% (based on 155, 156, and 163 substructures, respectively) in changing the number of peaks employed from four to six to eight. In the mass range 159-270 (MF4B) increasing the number of ions used for matching from five to eight increased the recall from 17.9% to 18.6% (based on 127 and 139 substructures, respectively). Those trends are not surprising in light of the increased uniqueness and scarcity of peaks in the higher mass ranges.[5]

Utility of High Mass Characteristic Ions

The use of eight peaks by MF4B (and also by MF4C) gave average recalls below that of MF4 which uses only five peaks between m/e 150 and $(M - 1)^+$. This can be due only in part to "confusion" caused by too much data used in

TABLE 10. DATA CLASSES GIVING HIGHEST RECALL VALUES FOR INDIVIDUAL SUBSTRUCTURES

Data class	Number of substructures with best recall[a]	Percentage of substructures with best recall
2A	58.3	33.0
(2A'[b])	(62.7)	(35.2)
3A	57.3	32.4
3B	36.5	20.6
4A	17.3	9.8
4B	4.3	2.4
4C	3.3	1.9

[a] The averaged STIRS results for each substructure were examined to find which data class gave the highest recall value; fractional credit was given for ties.

[b] The STIRS data for data class 2A' was ignored in determining the other results of the Table. The 2A' results shown were then determined in the same way, ignoring the 2A data.

matching, as (vide supra) increasing the number of peaks used by MF4B from five to eight gave a small increase in average recall. Thus we will extend (Table 11) the mass range of MF4B to $(M - 1)^+$, making the specifications for data class 4B little changed from those of the original class 4. The number of possible molecular fragments which can produce a particular mass increases rapidly with increasing mass, apparently offsetting the higher uniqueness of these peaks, so that a much broader mass range is advantageous at higher *m/e* values.

Requirement of Even-Mass and Odd-Mass Ions

For the lowest mass range (MF 2) equal numbers of the largest even- and odd-mass ions were originally chosen[10] to insure that both odd-electron and even-electron ions were included, as these are produced by different mechanisms. On average, however, there is little difference in the results from the use of the largest four even- and four odd-mass ions in the range *m/e* 6-88 (MF2A) and the use of the eight largest peaks in the same range (MF2A'). MF2A gave an average of 95.5% precision and 32.7% recall for 174 of the 204 substructures, while MF2A' gave an average of 97.3% precision and 33.4% recall for 171 substructures. Some individual substructures show substantial differences in recall between MF2A and MF2A', but this was not felt to be a sufficient justification for the use of both data classes in view of the additional matching time and storage required to retain both. Because the overlapping data class 3A uses only the most abundant ions, it was decided to use 2A instead of 2A', thereby having low mass data classes which both require, and do not require, even and odd masses. A study of 20 commonly occurring substructures using MF2a and MF2a' produced results using 10 ions similar to the above eight ion results. The recall average of 28.2 and 31.1%, respectively, for MF2a and MF2a' with 90% precision follow the trend above. The difference in recall can, in part, be accounted for by the fact that often there may not be five odd-electron ions of importance in the mass range.

Overlapping Mass Ranges

Evaluation of the effects of overlapping mass ranges is possible by comparison of the results for data classes 3A, 4A, and 4C to the results for the adjacent ranges 2A, 3B, and 4B. Enumeration for each data class of the number of substructures for which a maximum recall was achieved (Table 10) shows that such overlapping is of substantial value; predictions by MF3A and MF4A have 32 and 10%, respectively, of such substructures. The recall average (Table 2) of 34.6% for MF3A is the highest average of all the individual characteristic ion data classes, and for 64 substructures the recall performance of MF3A was greater than or equal to that for MF2A, MF2A', or MF3B. In contrast to the use of ten instead of eight peaks, the additional information obtained from characteristic ions by the use of overlapping mass ranges appears to be more than sufficient to justify the extra storage and time required for their use. The fact that nearly two-thirds of the substructures are found with maximum recall by MF2A and MF3A suggests that an additional data class in this mass range would be helpful. We propose to add one covering *m/e* 47-102 (data class 2B), in part compensating for the extra computer time and storage requirements through reductions in the number of peaks used in other data classes (Table 11).

TABLE 11. RECOMMENDED DATA CLASSES FOR CHARACTERISTIC IONS

Data class[a]	Maximum number of peaks	Mass range
2A	4 even-mass, 4 odd-mass	6 - 88
2B	8	47 - 102
3A	7	61 - 116
3B	7	89 - 158
4A	6	117 - 200
4B	6	159 - (M - 1)
11.1[b]	2A + 2B + 3A + 3B + 4A + 4B	

[a] Note that classes 2A, 3B, and 4B correspond closely to the original (reference 3) data classes 2, 3, and 4, respectively, except for the increased number of peaks.

[b] To replace the original data class 10.

Combination Match Factors

It has been shown (vide supra) that an arithmetic combination of the MF 1 - MF 6 match factor values (the "overall match factor," MF 11) give significantly higher average values of both recall and precision for substructure identification than those found for the individual data classes. Two combinations of characteristic ion match factors were used in this study, one employing MF2A plus MF3A through MF4C (MF11.1), the other MF2A' plus MF3A through MF4C (MF11.2). As expected from the MF2A and MF2A' results discussed above, there is no significant difference between the results for these two different combinations (Table 9), and only the MF11.1 results will be discussed. As found for MF 11, combining individual MF values substantially improves the precision of the results, bringing the performance up to the predicted 98% confidence level. The recall value of 47% for MF11.1 is far superior to that of any of the individual characteristic ion data classes, and this should increase with any improvements resulting from modifications of the data classes recommended in Table 11.

The average recall value of 47% for MF11.1 is substantially higher than the 40% value for MF 10, which uses two peaks every 14 mass units from *m/e* 90 to M^{+}. Therefore it is recommended that MF11.1 be used in place of MF 10 in STIRS; this will effect a substantial saving in the bulk data storage requirement, as the MF 10 data must be stored in variable length records.

Although the original version of STIRS did not have a match factor combining MF 2, 3, and 4 to which MF11.1 can be directly compared, the average recall using MF11.1 actually approaches that for MF 11, and the 183 substructures found with non-zero recall is a slight increase over the number for MF 11. The substructures for which the MF11.1 recall values were substantially improved over those of MF 11 are shown in Table 12. One-third of those substructures for which MF 11 had zero recall can be identified in some cases by MF11.1. The substructures benzphenanthrene-3,4- and indazole were identified correctly by MF11.1 for every compound in the file (9 and 18 compounds, respectively) containing the substructure. Because the MF 11 calculation includes neutral loss and ion series contributions, its performance should be substantially improved when the new data classes (MF2A - 4B) are included. The use of arithmetic combinations of data classes, such as MF 11 and 11.1, for matching takes no disk storage and little additional calculation time, yet appears to provide the most important and reliable structural information of all the STIRS results.

Examples

A statistically valid comparison of STIRS and manual interpretive methods was not attempted, as it would hardly be feasible for a human interpreter to examine several hundred spectra to predict the presence of each of 204 substructures. STIRS has been designed to be an aid to, not a replacement for, the interpreter. For STIRS predictions of substructures, it should be kept in mind that a 99% probability of presence is also a 1% probability that the substructure is absent; because there are several hundred acceptable (>94% precision, >30% of MF 11 or MF11.1 recall) combinations of the 13 match factors and 204 substructures for which STIRS makes predictions, this

TABLE 12. SUBSTRUCTURES FOR WHICH MF11.1 HAS SUBSTANTIALLY SUPERIOR RECALL

WLN	Substructure	Recall, %	
		MF11	MF11.1
ZR BV	anthraniloyl	0	25
MVR	benzamido	0	10
SHR	benzenethiol	0	8
T B656 EN HMJ	carboline, beta	0	60
T B656 HOJ	dibenzofuran	0	60
T66 CNJ	isoquinoline	0	11
QR B1U	salicylidene	0	8
SWQ	sulfonic acid	0	14
T56 BN DOJ	benzoxazole	55	75
L C666J	anthracene	14	43
L C6 B666J	benzphenanthrene-3,4-	88	100
T56 BN DSJ	benzothiazole	36	70
V1U2	crotonyl	17	48
T56 BMNJ	indazole	50	100
T56 BMT&J	indoline	33	66
T56 ANJ	indolizine	63	83
T5NOJ	isoxazole	43	66
V1V	malonyl	23	43
T C666 BN INJ	phenazine	47	67
T C666 BM ISJ	phenothiazine	60	90
T56 BVMVJ	phthalimide	37	63
T5MTJ	pyrrolidine	20	40
1U1	vinyl	23	46

probability will, on average, produce several predictions which are incorrect, at least on a strict chemical basis. Thus the interpreter must evaluate the STIRS substructure predictions in light of the probable significance of the other STIRS results, other mass spectral data, and other information available on the unknown. The incorrect prediction of a particular substructure often occurs because the mass spectral data typical of the substructure resemble those of the correct answer; thus a mass spectrometrist would not be surprised if the STIRS results for the low mass characteristic ions (MF2A) indicated phenyl for a pyridine compound, even though a chemist would call this substructure an incorrect answer. Perhaps the aid to interpretation supplied by the improved characteristic ion data classes can best be illustrated by a few "unknown" examples.

The compound 1-undecanol (WLN, Q11), when analyzed by STIRS, gives confidence values of greater than 99% for the presence of two substructures: alkyl chain ≥3 carbons (by MF3B, MF11.1 and MF 11) and hydroxyl (Q, by MF 11). Incorrect indications of S and SH by MF2A and MF11.1, respectively, are not entirely misleading, as for these data classes such compounds show behavior similar to that of alcohols. The other incorrect predictions are cyclopentyl by MF4A and cyclohexanone by MF2A. The top 15 compounds retrieved by MF11.1, which uses only characteristic ions, are five alkan-1-ols, three of the corresponding thiols, and five _n_-alk-1-enes. The mass spectra of alkan--1-ols are characterized by an initial loss of water, making their characteristic ion data similar to those from the spectra of alk-1-enes. However, the top 15 compounds of the MF 11 results, which in addition use information on neutral losses, are all alcohols, eight of which are _n_-alkan-1-ols. Thus the STIRS results should substantially help the interpreter in obtaining the structure of the unknown.

For the compound _o_-hydroxyphenyl _tert_-butyl sulfone (WLN, QR BSWX) STIRS indicated at the >99% confidence level the following substructures: hydroxyl (Q, by MF3A, MF11.1), phenyl (R, by MF3A, MF11.1 and MF11), hydroxyphenyl (QR, by MF11.1 and MF11), sulfonyl (SW, by MF3A, MF3B, MF4A, MF11.1 and MF11), sulfur (S, by MF11.1 and MF11), and double branch (X, by MF2A). Incorrect indications of ester (VO, by MF8), linking oxygen (O, by MF7 and MF11), benzoate (OVR, by MF11), and salicyloyl (QR BV, by MF1 and MF11) result from the similar mass spectral behavior of a molecular fragment such as _o_-HOC_6H_4-CO_2-, for which these substructures would be indicative, and the fragment _o_-HOC_6H_4-SO_2- of the correct structure. The indication of cyclopropyl (L3TJ, by MF4B) is incorrect and unrelated to the actual structure. The top 8 compounds retrieved by MF11.1 and the top 4 retrieved by MF 11 are 2-hydroxyphenylsulfones. Thus in this example the characteristic ion data classes, which are combined in MF11.1, actually give the best indication of structure. It was previously found that the overall match factor, MF 11, gives the most reliable substructure predictions; it appears that MF11.1 is also a much better indicator than the individual match factors.

For the compound methyl 8-phenylnonanoate (WLN, 1YR&6VO1) STIRS gives substructure assignments at the >99% confidence level for: alkyl chain ≥3 carbons (by MF3B, MF11.1 and MF11), phenyl (R, by MF3A, MF11.1 and MF11), and single branch (Y, by MF3A and MF11.1). Indications of acrylyl

(V1U1, by MF4B) and the 6.7-dehydro steroid skeleton (L E5 B666 LUTJ, by MF4A) are errors from 99% confidence level predictions. Surprisingly, the only high confidence prediction of the ester function came from MF4C, which is being dropped (vide supra) because of generally poor recalls. In the MF11.1 and MF 11 results, nine and five, respectively, of the top 15 compounds are 2-phenylalkanes, and two and four, respectively, are methyl esters of phenylnonanoic acid. Thus the results from the characteristic ions (MF11.1) nicely complement those from the overall match factor (MF 11); by using the STIRS results and deducing the molecular weight the interpreter should easily be able to obtain at least a close approximation to the molecular structure.

Improvement of Other STIRS Data Classes

The substantial increases in recall for substructures and the valuable information obtained from the use of the new overall match factor MF11.1 suggest that such modifications could also be used to improve the performance of other STIRS data classes, again applying the statistical methods to evaluate the success of such modifications. At present we are studying the effect of adding overlapping mass ranges and a special overall match factor to the neutral losses data class.

Implementation of STIRS Improvements

The substructure and data class improvements have been implemented and are operational on our laboratory PDP-11/45 computer, and thus could of course be made available over the long-distance phone link to outside users. This is made possible in direct data transmission by a program written by Dr. Walter M. Shackelford of ERL, Athens. For obvious reasons we would rather have outside users use either of the generally available STIRS systems on the Cornell IBM-370/168 over TYMNET, or on the NIH PDP-10 computer; for the latter contact Dr. G. W. A. Milne of NIH or Dr. S. R. Heller of EPA, Washington, DC. We will take the responsibility of implementing the substructure and data class improvements on the Cornell IBM-370/TYMNET system, however, early in 1976. Note that in particular EPA laboratories the PDP-8 computers used for data acquisition and reduction on the GC/MS systems now have the capability of direct transmission of unknown mass spectra over the TYMNET phone link to the Cornell PBM/STIRS system. This has the added advantage that for at least the near future the Cornell PBM/STIRS system will have reference spectra of a few thousand more compounds than any other available system.

Testing of PBM/STIRS by EPA

As emphasized in the recommendations, we feel that it is highly important that the combined PBM/STIRS system be tested on real unknowns by qualified personnel over an extensive period. Initial testing of the STIRS system on the NIH PDP-10 has been extremely disappointing for a variety of reasons. Poor communications make it difficult to resolve problems of user education. There appear to be misunderstandings concerning the "interpretive" nature of STIRS; all of the top compounds selected in each data class should be examined for structural consistencies, and one (or more) trial molecular weight(s) should be entered, even if it is just an educated guess. Inordinately long computer times (~15 minutes per spectrum) are required; these presumably give costs of $200 or more per spectrum, which severely lower the probability that STIRS will prove useful in a test on this computer. It is strongly recommended that the Cornell PBM/STIRS system on TYMNET be used instead; PBM has been designed specifically as a prefilter to STIRS, so that the two together are a much more effective system than either separately. Certainly the improved version of STIRS described above should be used in this extensive test; for example, note the nearly "perfect" information precision and recall values for the carcinogenic substructure benzphenanthrene-3,4- (Table 12). Further, there are full-time personnel in Cornell's Office of Computer Services who are interested in insuring that these systems are working properly, and that the user receives the information he needs. The wide acceptance which the Cornell system has already received from experienced mass spectrometrists is a strong indication of the value of PBM/STIRS.

As an alternative, if a matching system such as PBM could be implemented in real time on the GC/MS computers in the EPA laboratories, this important prefiltering would then be done, and only spectra unidentified by PBM could then be submitted to STIRS, as has been outlined in Section II.

SECTION VII

REFERENCES

1. Pesyna, G. M., and F. W. McLafferty. Computerized Structure Retrieval and Interpretation of Mass Spectra. In: Determination of Organic Structures by Physical Methods, Nachod, F. C., J. J. Zuckerman, and E. W. Randall (Eds.). New York City, Academic Press, 1975. Volume 6.

2. Salton, G. Automatic Information Organization and Retrieval. New York, McGraw-Hill, 1968.

3. McLafferty, F. W., R. H. Hertel, and R. D. Villwock. Probability Based Matching of Mass Spectra. Rapid Identification of Specific Compounds in Mixtures. Org. Mass Spectrom. 9:690, 1974.

4. Abramson, F. P. Anal. Chem. 47:45, 1975. See also: Abramson, F. P. Proc. of the 21st Ann. Conf. on Mass Spec. and Allied Topics. San Francisco, ASMS, 1973. p. 76; Abramson, F. P., and M. F. Schulman. Proc. of the 22nd Ann. Conf. on Mass Spectrom. and Allied Topics. Philadelphia, ASMS, 1974. p. 453.

5. Pesyna, G. M., F. W. McLafferty, R. Venkataraghavan, and H. E. Dayringer. Statistical Occurrence of Mass and Abundance Values in Mass Spectra. Anal. Chem. 47:1161, 1975.

6. Stenhagen, E., S. Abrahamsson, and F. W. McLafferty. Registry of Mass Spectral Data. New York City, Wiley-Interscience, 1974.

7. Freund, J. E. Mathematical Statistics. Englewood Cliffs, Prentice-Hall, 1962. p. 46.

8. Winer, B. J. Statistical Principles in Experimental Design. Second Edition. New York City, McGraw-Hill, 1972.

9. Siegel, S. Non-Parametric Statistics. New York City, McGraw-Hill, 1956. Chapter 2.

10. Kwok, K.-S., R. Venkataraghavan, and F. W. McLafferty. Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System. J. Amer. Chem. Soc. 95:4185, 1973.

11. Hertz, H. S., R. A. Hites, and K. Biemann. Anal. Chem. 40:681, 1971.

12. Pesyna, G. M. Computerized Structure Retrieval and Interpretation of Mass Spectra: The Design and Evaluation of a Probability Based Matching System Using a Large Data Base. Ph.D. Thesis. Ithaca, Cornell University, 1975. 269 p.

13. Meyerson, S. and E. K. Field. J. Chem. Soc. (B). 1001, 1966.

14. Costello, C. E., H. S. Hertz, T. Sakai, and K. Biemann. Clin. Chem. 20:255, 1974.

15. Dayringer, H. E. Computer-Aided Interpretation of Mass Spectra: An Improved STIRS Program Giving Information on Substructure Probabilities. Ph.D. Thesis. Ithaca, Cornell University, 1976. 162 p.

16. Kent. P., and T. Gäumann. Helv. Chim. Acta. 58:787, 1975.

17. McLafferty, F. W., M. A. Busch, K.-S. Kwok, B. A. Meyer, G. Pesyna, R. C. Platt, I. Sakai, J. W. Serum, A. Tatematsu, R. Venkataraghavan, and R. G. Werth. A Self-Training Interpretive and Retrieval System for Mass Spectra. The Data Base. In: Mass Spectrometry and NMR Spectroscopy in Pesticide Chemistry, Biros, F. J., and R. Haque (Eds.). New York, Plenum Press, 1974. p. 49.

18. Isenhour, T. L., B. R. Kowalski, and P. C. Jurs. Critical Review Anal. Chem. 4:1, 1974.

19. Justice, J. B., and T. L. Isenhour. Anal. Chem. 46:223, 1974.

20. Tunnicliff, D. D., and P. A. Wadsworth. Anal. Chem. 45:12, 1973.

21. Franzen, J., and H. Hillig. Adv. Mass Spectrom. 6:991, 1974.

22. Chemical Substructure Dictionary, Institute for Scientific Information, Philadelphia, 1974.

SECTION VIII

LIST OF PUBLICATIONS

References 1, 3, 5, 12, 15, and 17 are publications which have resulted from this research grant. In addition, the following articles have either been published or submitted for publication:

McLafferty, F. W., R. Venkataraghavan, K.-S. Kwok, and G. Pesyna. A Self-Training Interpretive and Retrieval System for Mass Spectra. Adv. Mass Spectrom. **6**:999, 1974.

Dayringer, H. E., G. M. Pesyna, R. Venkataraghavan, and F. W. McLafferty. Computer-Aided Interpretation of Mass Spectra. Information on Substructural Probabilities from STIRS. Org. Mass Spectrom. (accepted).

Venkataraghavan, R., G. M. Pesyna, and F. W. McLafferty. Computer Identification and Interpretation of Unknown Mass Spectra Utilizing a Computer Network System. In: Computer Networks and Chemistry, Lykos, P. (Ed.). Washington, American Chemical Society, 1975. p. 183.

Pesyna, G. M., R. Venkataraghavan, H. E. Dayringer, and F. W. McLafferty. A Probability Based Matching System Using A Large Collection of Reference Mass Spectra. Anal. Chem. (submitted).

Dayringer, H. E., and F. W. McLafferty. Computer-Aided Interpretation of Mass Spectra. Increased Information From Characteristic Ions. Org. Mass Spectrom. (submitted).

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/4-76-046	2.	3. RECIPIENT'S ACCESSION NO.

4. TITLE AND SUBTITLE Computer Interpretation of Pollutant Mass Spectra	5. REPORT DATE October 1976 (Issuing date)
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Fred W. McLafferty	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Cornell University Department of Chemistry Ithaca, NY 14853	10. PROGRAM ELEMENT NO. 1BA027
	11. CONTRACT/GRANT NO. R-801106
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Athens, Georgia 30601	13. TYPE OF REPORT AND PERIOD COVERED Final 10/1/72 - 9/30/75
	14. SPONSORING AGENCY CODE EPA-ORD

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The objective of this research was to improve systems for computer examination of the mass spectra of unknown pollutants. For this we have developed a new probability based matching (PBM) system for the retrieval of mass spectra from a large data base, and have substantially improved the interpretation of unknown mass spectra using the self-training interpretive and retrieval system (STIRS). PBM was designed as a prefilter to STIRS; if an unknown mass spectrum can be identified with a sufficiently high confidence by PBM, interpretation of the spectrum using STIRS is not necessary. The PBM system provides more efficient retrieval than presently accepted systems; it incorporates a "reverse search" algorithm, and through the use of weighted mass and abundance data provides a statistically valid prediction of the confidence of the matches found. STIRS has been improved to give a confidence-level prediction of the presence of ~200 particular substructural features in the unknown molecule. Extensive studies have been made to improve the data selection for most data classes used by STIRS, resulting in a much higher level of overall system performance. Operation efficiencies of both PBM and STIRS have been improved dramatically so that both require less than 1 minute on a laboratory PDP-11/45 computer. STIRS has been made available for outside use by long-distance phone connections to this PDP-11/45, and recently both PBM and STIRS have been made operational on the Cornell IBM-370/168 so that these are available internationally over the TYMNET computer network system.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Computer programming Mass spectra Algorithms Information retrieval Chemical analysis		07B

18. DISTRIBUTION STATEMENT Release Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 67
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

☆USGPO: 1976 — 757-056/5421 Region 5-11