Retrieval and Interpretative Computer Programs for Mass Spectrometry

FRED W. MCLAFFERTY* and DOUGLAS B. STAUFFER

Chemistry Department, Cornell University, Ithaca, New York 14853-1301

Received April 8, 1985

Using the modern gas chromatograph/mass spectrometer (GC/MS), an interpreter may be faced with more than 100 unknown electron-ionization mass spectra per hour. The Probability Based Matching (PBM) program yields real-time identifications for such a GC/MS output. Because GC separation can be incomplete, PBM employs reverse searching for improved identification of mixture components. Forward searching, which is more specific for pure samples, is also automatically incorporated by matching the residual spectra obtained by subtracting the best matching reference spectra from the unknown. The 81 000 different spectra of 68 000 different compounds of the Wiley/NBS data base were measured under a wide variety of experimental conditions; to compensate for this variability, PBM employs peak "flagging" and abundance "scaling". These and other values reflecting the degree of match are converted statistically into a single "reliability" value directly indicating the probability that the structure prediction is correct. With these improvements the first answer retrieved for pure and 60% mixture components was correct, or difficult to distinguish from the correct answer by mass spectrometry, in 97% and 93% of cases, respectively. If the unknown is not represented in the reference file, the Cornell Self-Training Interpretative and Retrieval System predicts its molecular weight, number of chlorine and bromine atoms, and substructural features present. For 589 substructures, a quantitative "reliability" value is assigned to the STIRS prediction.

INTRODUCTION

The application of compound identification techniques to complex mixture problems has increased dramatically in the last decade. Pollutants, overdose victims, insect chemical communication, body fluid diagnoses, forensic evidence, drug metabolites, and chemical taxonomy are all problems that have greatly benefited from the increased sensitivity, specificity, and speed of analytical instrumentation. Mass spectrometry (MS) is by far the most widely used technique for such identifications of organic compounds, especially when coupled to an efficient separation device such as the gas or liquid chromatograph (GC, LC).[1] The modern GC/MS instrument can produce characteristic electron-ionization (EI) mass spectra of more than a 100 subnanogram components from a complex mixture per hour. Even several mass spectrometrists would be challenged to keep up with the interpretation required by this output, a problem that has led to a variety of computer programs to aid in this task.[2-14] Because the algorithms that appear to exhibit the best performance[9-12,14,15] as well as widest use[16] were developed at Cornell,[6,7,9] these will be the main focus of this review.

Identification algorithms can be classified in two main categories, "matching" (or retrieval) and "interpretative".[10] In general, matching programs require less human intervention and so are used first. If a sufficiently good match is not retrieved from the reference file, an interpretative program can then be used to predict structural information. The Cornell matching and interpretive algorithms for EI mass spectra are respectively Probability Based Matching (PBM)[7,9] and the Self-Training Interpretative and Retrieval System (STIRS).[6]

These algorithms have been designed for the identification problem in its broadest sense—the "total unknown". If other information restricts the unknown structure to a narrower area of compound types, the chemist may feel that it would be wasteful to search for the structure throughout the universe of possible compounds. However, artifacts are always possible, such as phthalates in a blood sample collected with plastic tubing. Further, if the identification program is sufficiently fast to propose structures before the end of the GC/MS run without human effort, this information will be valuable at least for confirmation purposes. Algorithms designed for very specific compound types, such as the estrogens,[5] should give more accurate and detailed information where they are applicable. However, substantial peformance improvements can also be achieved with "total unknown" programs by adding to the data base reference spectra of the most commonly encountered compounds measured under the same experimental conditions used for unknowns.

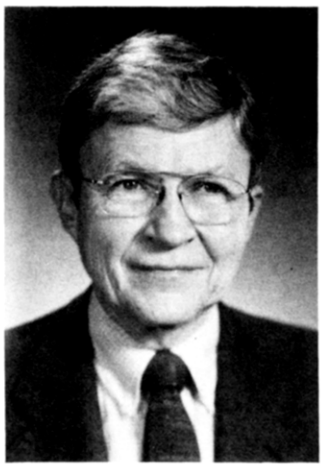
Finally, it should be emphasized that such algorithms should be viewed only as aids to, not as replacements for, the interpreter. The computer can only use the spectral data and correlations it has been given, while the interpreter can remeasure the unknown data under more rigorous conditions, search for new correlations, or run a reference spectrum of the predicted compound under the same experimental conditions used for the unknown. The interpreter also has the ultimate responsibility for understanding the basic limitations of the analytical method employed—there is yet no "universal analyzer".

EXPERIMENTAL PROCEDURES

Versions of both PBM and STIRS have been implemented on a variety of computer systems (H/P-1000, DEC-11/45, IBM-370/168 and 4341, Data General), are available on commercial GC/MS systems such as those of Hewlett-Packard and Nicolet, and can be accessed directly on the Cornell computer[17] through TYMNET and TELENET networking systems. Earlier versions of the software are available from the Quantum Chemistry Program Exchange, University of Indiana, and the current version from the authors. The 1983 Wiley/NBS Registry of Mass Spectral Data,[18] containing >81 000 different mass spectra of >68 000 different compounds, and its earlier versions were used as reference files, eliminating ~3000 spectra of isotopically labeled compounds.

Modifications to the algorithms were tested using "recall-reliability" plots[19] similar to evaluation methods developed for algorithms to retrieve references from libraries.[20] Unknown spectra were chosen at random from the data base and then excluded from the reference file used in the test. The approximately 400 "unknown" spectra chosen for PBM testing were those of compounds that are represented by one or more other spectra in the data base. For STIRS, approximately 900 "unknown" test spectra were used. In both cases, these numbers of test spectra were shown to be sufficient by running a

0095-2338/85/1625-0245$01.50/0 © 1985 American Chemical Society



Professor Fred W. McLafferty came to Cornell in 1968 after holding positions that included Director of Dow Chemical's Eastern Research Laboratory for basic research and professor at Purdue University. He is a member of the U.S. National Academy of Sciences and an honorary member of the Italian Chemical Society. He has been chairman of the ACS Analytical Division and the AAAS Chemistry Section and is a member of the NRC Board on Chemical Sciences and Technology and the Pimentel Committee to Survey Opportunities in the Chemical Sciences.

Douglas Brent Stauffer was born July 5, 1957, in Lancaster, PA. In May 1979 he graduated magna cum laude from Grace College, Winona Lake, IN, with a B.A. in Mathematics, Chemistry, and Biology. He earned his Ph.D. degree in August 1985 at Cornell University, where he is now a Postdoctoral Research Associate. He is a member of the American Chemical Society, the American Society for Mass Spectrometry, Sigma Xi, and the Association for Computing Machinery.

different random test set of the same size, producing results within ~2% of those of the first set.

PROBABILITY-BASED MATCHING

The Data Base. A matching algorithm is fundamentally limited by the quality and quantity of the reference spectra. To be used with a "total unknown", these criteria are to some extent mutually exclusive. Optimum quality demands that each reference spectrum, as well as the unknown, be run under the same experimental conditions, but few laboratories control conditions to a common standard to make possible the exchange of reference spectra. Our approach has been to emphasize quantity, collecting as many spectra as possible, irrespective of instrument type or experimental conditions. Quality standards are only imposed in retaining multiple spectra of the same compound. The current data-base[18] size results from a recent near doubling at Cornell, and a similar expansion is under way here now.

Because the reference spectra of the data base were measured under such a wide variety of experimental conditions and standards, it is imperative that the matching program contains features designed to compensate for the resulting disadvantages. Two features of PBM designed for this purpose, "flagging"[9] and "scaling",[21] will be discussed below. In addition, the spectrum subtraction procedure[22] adding forward search capabilities (vida infra) also appears to have a beneficial compensatory effect.[15,23] Finally, probably the most important aid for this problem is the inclusion in the data base of different spectra of the same compound. It has been demonstrated that this increases substantially the accuracy of PBM structure prediction (e.g., a 33% reduction in incorrect first retrievals in a comprehensive trial).[24]

Reverse vs. Forward Searching. As shown by ourselves[2,7] and by Abramson,[25] the ability of an algorithm to match two or more components in the mass spectrum of a mixture is greatly aided by only requiring the peaks of the reference spectrum to be in the unknown, not vice versa, termed "reverse searching". Thus, the presence in the unknown spectrum of peaks not in the reference (such as those from another component) does not lower the degree of match as would be the case with a forward search. However, for an unknown shown separately to be a pure compound, forward searching has the "no-band" advantage[2,26] that reference spectra which do not contain peaks of the unknown can be eliminated as possible matches. For the mass spectra of pure compounds, the early PBM system exhibited retrieval capabilities comparable[9] to those of the Biemann-MIT forward-searching program,[4] while its reverse-searching capabilities, as expected, were substantially superior for unknown spectra of mixtures. Because even chromatographic peaks from capillary GC columns have a substantial probability of containing more than one component,[27] a reverse-searching algorithm should be preferable for most GC/MS uses. A new algorithm[15,23] combining both reverse- and foward-searching capabilities, showing improved performance for unknown spectra of both mixtures and pure compounds, is described below.

Data Weighting. PBM was modeled directly on the sophisticated reference retrieval systems that have been developed for libraries. For these, "it is customary to identify document or query content by sets of terms...with term weights to reflect the presumed importance of each term.[28] Mass spectra contain two main types of terms, abundance and mass (m/z, mass to charge ratio). The weights of these terms are determined directly from their probability of occurrence in the data base.[7,29] Abundances closely follow a log normal distribution, as first pointed out by Grotch;[30] for this, PBM uses $\log_2$ "A" values in which $A = -1$ corresponds to abundances of 0.24%–<1%, 0 to 1%–<3.4%, 1 to 3.4%–<9%, 2 to 9%–<19%, 3 to 19%–<38%, 4 to 38%–<73%, and 5 to 73%–100% (i.e., the number of spectra having abundance values ≥73% is 2^{-5} of those having values ≥1%). For mass value weights, PBM uses $\log_2$ "uniqueness" (U) values, based on the proportion of compounds in the data base whose spectra show a peak of ≥1% relative abundance at that specific mass. The quantitative validity of this weighting approach, already well established for document retrieval from libraries,[20] has been completely confirmed by the PBM application to mass spectral retrieval.[9] Qualitatively, this can be seen by considering that peaks such as m/z 41 (e.g., $C_3H_5^+$) and 57 (e.g., $C_4H_9^+$, $C_3H_5O^+$) are very common in mass spectra, while those such as m/z 47 (e.g., CH_3S^+) and higher mass peaks occur in far fewer reference spectra. Thus, if both the unknown and a reference have matching m/z 41 and 57 peaks, this is much poorer evidence

that both originate from the same compound than if they both have matching m/z 47 and 441 peaks.

Data-Base Variability. There are two important quality deficiencies of the Wiley/NBS data base for which the algorithm must be specially designed. Errors are common, such as mass values measured incorrectly or artifact peaks due to impurities or background. Second, all GC/MS systems are subject to mass discrimination, so that relative abundance values for individual mass peaks can differ by more than 1 order of magnitude. To compensate for these two problems, peak flagging and scaling, respectively, have been incorporated into PBM. For each reference spectrum only the 15–26 (depending on molecular weight) most important peaks (highest $U + A$ values) are used in matching.

If for every peak in a particular such "condensed" reference spectrum there is a corresponding peak in the unknown, with the abundance of each unknown peak just 50% of that of the reference, this would indicate that a component of the unknown present at approximately 50% concentration is an excellent match for the reference. To ascertain this information, PBM first determines for each reference m/z value the ratio ρ of the unknown abundance to the reference abundance. If the reference compound is present in the unknown, there should be no zero values (within experimental error) for ρ, and the minimum ρ value should be indicative of the proportion of reference present in the unknown (note that contributions from other components to an unknown peak will increase its ρ value). However, if a correct reference spectrum (that of a component of the unknown) contains an artifact or mass-error peak not in the unknown, the ρ value for this mass would be zero, incorrectly indicating that the reference does not match. To avoid this, after the first match PBM sets a higher ρ value, "flags" (eliminates) all mass peaks below this limit, and repeats the matching. The flagging and rematching process is repeated twice more, and the highest reliability value of the four matches is saved. Tests[9,21] showed that a fourth flagging operation is counterproductive by increasing the probability of incorrect matches. Obviously, removal of errors from the data base, such as those pointed out by the necessity of flagging, should reduce the desirable number of flagging operations.

"Scaling" compensates for abundance variations that are mass dependent, such as instrumental mass discrimination and variations in sample pressure during scanning of the mass spectrum.[21,31] For the former, many instruments transmit high-mass ions less efficiently than low-mass ions, and for quadrupoles, even a minimum in relative mass discrimination with increasing mass has been observed. Sample pressure in the ion source can go through a maximum for spectrum scanning as a GC peak top is eluted or as direct-probe sample heating exhausts the sample. Linearly tilting these unknown spectra to optimize the match gave considerable improvement in PBM performance.[21] A further improvement was effected by quadratic fitting, which allows the discrimination function to go through one maximum or minimum.[21]

Reliability Ranking. For most retrieval systems the value scale indicating the match ranking (e.g., Similarity Index,[4] Confidence Value[7]) has little quantitative relationship to the probability that the correct answer has been retrieved, although this can be inferred by the user from such values with experience. A special danger to the inexperienced user is to conclude that the best matching spectrum must have a high probability of representing the correct compound, forgetting that this compound might not actually be in the reference file. To convert the degree of PBM matching into an actual "reliability" value, statistical evaluations were made[21] on the probability of retrieving a correct answer vs. the values of four matching parameters: the original confidence value (K) based on the $U + A$ values of matching peaks, the presence or absence of a match for the reference molecular ion, the total $U + A$ value of all peaks flagged, and the degree of scaling used. Using the resulting reliability values to retrieve reference spectra gave a substantial improvement in PBM ranking;[21] these represent spectral data in addition to the U and A values whose proper weighting affects performance. The reported reliability values were further modified to reflect these improvements.

Table I. Results from PBM without and with Added Forward Searching for the Unknown Spectrum of Pure 1,1-Diethoxypropane

compounds retrieved	reliability (%) reverse search	+forward search	new rank
propanal	50	40	8
1,1-diethoxypropane	48	85	1
3-hydroxy-3-methyl-2-butanone	48	10	20
2,2-diethoxyethanol	45	58	7
3-pentanol	44	45	11
2-(1,1-dimethylethoxy)ethanol	44	40	12
3-pentanol	44	50	9
3-pentanol	44	35	13
diethoxymethane	43	60	6
diethoxymethane	42	62	5
ethyl thiocyanate	42	66	3
1-pentylhydroperoxide	42	35	14
(ethoxymethyl)oxirane	38	65	4
propanal	38	14	19
bis(1-methyl-2-hydroxyethyl) ether	38	50	10
3-pentanol	37	35	15
2-(2-hydroxypropoxy)-1-propanol	37	25	17
morpholine	35	35	16
1,1-diethoxypropane	34	84	2

Forward-Searching Capabilities Added to the PBM Reverse-Search System. The PBM results include a "percent contamination" ($\%C$) value based on the relative proportion of the unknown peaks used in matching the reference spectrum. This, in effect, is "no-band" information,[2,26] indicating the degree to which the unknown spectrum contains peaks not in the reference. For unknown mass spectra of *pure* compounds, inclusion of this additional weighting factor in the reliability calculations gave a substantial increase in PBM performance.[15,23]

For most unknowns, however, the user finds the necessity of assessing unknown purity inconvenient, and of course, this could even be inaccurate. To avoid this problem, from the unknown mass spectrum the algorithm attempts to generate spectra of the individual components by subtracting each best matching reference spectrum (the full spectrum) from the unknown spectrum.[22] PBM rematching of each of these residual spectra gave an impressive performance improvement (e.g., Table I). For 60% mixture components, two-thirds of the class IV (closely related compounds) wrong answers were eliminated, while for pure compounds such wrong answers were reduced by nearly half (only a few percent below the performance when such unknowns could be designated as pure compounds).[23]

Artifact Retrievals in Reverse Searching. PBM users have complained that retrieved references can represent, misleadingly, only a part of the molecule, such as the PBM matching of 3-octene (reliability 32%) against an unknown spectrum actually of dioctyl phthalate (Table II, pure footnote *a*).[15] This is because the octene spectrum actually is closely similar to a substantial part of the spectrum of dioctyl phthalate, as shown in Figure 1. If the user can designate this unknown as that of a pure compound so that $\%C$ weighting factor can be applied, the predicted reliability values of the correct answers are generally increased (Table II, pure, footnote *b*) and those of octenes and similar compounds are greatly reduced. However, with the spectrum-subtraction procedure, such a purity designation is not necessary, and a similar performance

Table II. PBM Matching (Class IV) of Mass Spectra of Pure Bis(2-Ethylhexyl) Phthalate and Its 3:1 Mixture with 3-Octene

compound	pure			mixture		
	a	*b*	*c*	*a*	*b*	*c*
bis(2-ethylhexyl) phthalate[d]	47[e]	50	61	56	22	45
diisooctyl phthalate	42	37	50	35	20	37
2-methyl-1-butene	34	<10	6	43	24	54
(*Z*)-3-octene	32	14	13	(47	27	97)[f]
3,4-dimethyl-1-pentene	31	<10	8	<10	<10	<10
bis(2-ethylhexyl) phthalate	29	44	77	40	24	55
all octenes[g]	25	14	9	55	24	83
(number of octenes retrieved)	(2)	(1)	(2)	(13)	(10)	(12)
all C_nH_{2n} compounds	24	14	9	49	23	64
(number of C_nH_{2n} compounds)	(14)	(1)	(3)	(18)	(16)	(17)

[a] Results using the previous reverse-search system.[21] [b] With forward search, using the percent contamination value to correct the reliability value, assuming that the unknown is the spectrum of a pure compound. [c] With forward-search contribution using multiple subtraction but with no knowledge of the purity of the unknown. [d] The unknown spectrum itself was not in the reference file used. [e] Reliability value (%). [f] The spectrum used as the 25% component of the unknown. [g] Average results for all retrieved reference spectra of C_8H_{16} olefins; the number retrieved is shown on the next line in parentheses.

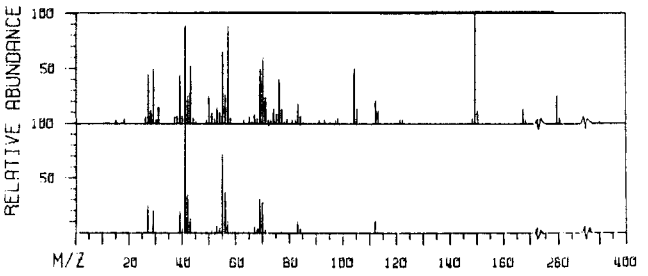


Figure 1. PBM matching of the upper mass spectrum, that of pure bis(2-ethylhexyl) phthalate, incorrectly predicted (Table I) the presence of (*Z*)-3-octene (bottom) with 32% reliability.

improvement is found (pure, footnote *c*). Further, if 3-octene is actually present in the unknown, it is recognized by the original reverse-search PBM (mixture, footnote *a*); performance is decreased by the incorrect assumption of purity (mixture, footnote *b*) but improved by the spectrum-subtraction procedure (mixture, footnote *c*).

Further Improvements. By use of the 76 663-spectra data base (omitting those of isotopically labeled compounds), for the 392 pure unknowns and the 130 60%-component unknowns, the first answer retrieved was correct in 84% and 79%, respectively, of the cases by class I criteria (same compound or a stereoisomer), and 97% and 93% by class IV criteria (compounds differing from the correct answer by structural changes listed as ones causing little change in the mass spectrum). Individual examination of the class IV wrong answers shows that possibly half of these differ structurally from the correct answer by additional structure classifications for which mass spectrometry is not particularly sensitive. Thus, we conclude that further improvements in matching this data base through modifications of the PBM algorithm itself will be marginal, and the primary efforts should be made in data improvement.

"Standard" Low-Resolution EI Reference Spectra. Project 44 of the American Petroleum Institute collected EI reference spectra whose abundances were highly reproducible (~±1%) between laboratories because they were measured on the same instrument (CEC 21-103) under conditions to give the same data for the *n*-butane spectrum. Budde[32] has proposed a similar system in which all spectra would be measured under instrumental conditions that give a standard set of relative abundances when measuring decafluorotriphenylphosphine. The Environmental Protection Agency has contracted to have

Table III. PBM Results[33] from High-Resolution Mass Spectrum of 3,5-Dimethylpyrazole[a]

compound	formula	reliability[b]
3,5-dimethylpyrazole	$C_5H_8N_2$	87
3,5-dimethylpyrazole	$C_5H_8N_2$	81
3,5-dimethylpyrazole	$C_5H_8N_2$	43
β-methylimidazole-4-ethanamine	$C_5H_8N_2$	43
2-ethylimidazole	$C_5H_8N_2$	43
(furfural)	($C_5H_4O_2$)	(42)
β-methylimidazole-4-ethanamine	$C_5H_8N_2$	41
(furfural)	($C_5H_4O_2$)	(40)
2,4-dimethylimidazole	$C_5H_8N_2$	36
(2-methyl-2,4-hexadiene)	(C_7H_{12})	(20)
(4-methyl-1,4-hexadiene)	(C_7H_{12})	(20)
1,2-dimethylimidazole	$C_5H_8N_2$	17
(4-hydroxypyridine)	(C_5H_5NO)	(15)
(4-pyrimidinone)	($C_4H_4N_2O$)	(15)

[a] Using the same uniqueness values as with normal PBM. [b] Class IV matches using normal PBM with unit mass resolution data; those eliminated by using ±0.003 mass tolerance are in parentheses.

Table IV. Molecular Ion Compositions of Compounds of Nominal Molecular Weight 140 in the Wiley/NBS Data Base

no. of compd	composition	Δ (millimass units)	required accuracy
1	$C_4H_{10}FO_2P$	0	
3	$C_6H_8N_2S$	0.6	235 000
13	C_8H_9Cl	−1.0	141 000
1	$C_8H_6F_2$	3.5	40 300
3	$F_5H_7F_3O$	4.7	30 000
1	$C_5H_6N_3O_2$	5.8	24 300
21	$C_7H_8O_3$	7.1	19 900
13	C_7H_8OS	−10.7	13 200
2	$C_7H_5FO_2$	−12.9	10 900
244	others	<116	<10 900
98	$C_{10}H_{20}$	116.3	1 200

several thousand reference spectra of common compounds measured in this way. Fortunately, the modern GC/MS computer system can also determine the relative mass discrimination of an instrument in measuring the decafluorotriphenylphosphine standard and apply this as a correction to the abundances in any reference or unknown spectrum measured under the same conditions. Modification of PBM to take advantage of these improved data will be relatively straightforward. The "window tolerances" allowed for abundance matching can be narrowed, and the weighting factor contributions to the final "reliability" value must be redetermined by using unknowns and reference spectra measured under the standard conditions. Then such an unknown can be matched first against the smaller standardized reference data with the new PBM parameters; if a suitable match is not found, the same unknown spectrum can then be matched with the current PBM program against the much larger data base.

High-Resolution Mass Spectra. Measuring the mass of a peak with sufficient accuracy determines its elemental composition, so that such "high-resolution" mass spectra obviously have a much higher information content and are more specific for structure elucidation. For example, if an unknown mass 43 peak can be shown to represent $C_2H_3O^+$ (43.0184), not $C_3H_7^+$ (43.0547), all hydrocarbon reference spectra can be eliminated as matching possibilities. A preliminary version of an "enhanced PBM" system has shown that utilization of such exact mass data substantially restricts the matching possibilities for the unknown spectrum of 3,5-dimethylpyrazole (Table III).[33]

The potential advantages in improved data weighting can be seen in the greatly increased uniqueness of a mass 140 peak whose composition can be shown to be $C_4H_{10}FO_2P$. Of 400 compounds of molecular weight 140 in the data base, only isopropyl methylphosphonofluoridate has this composition

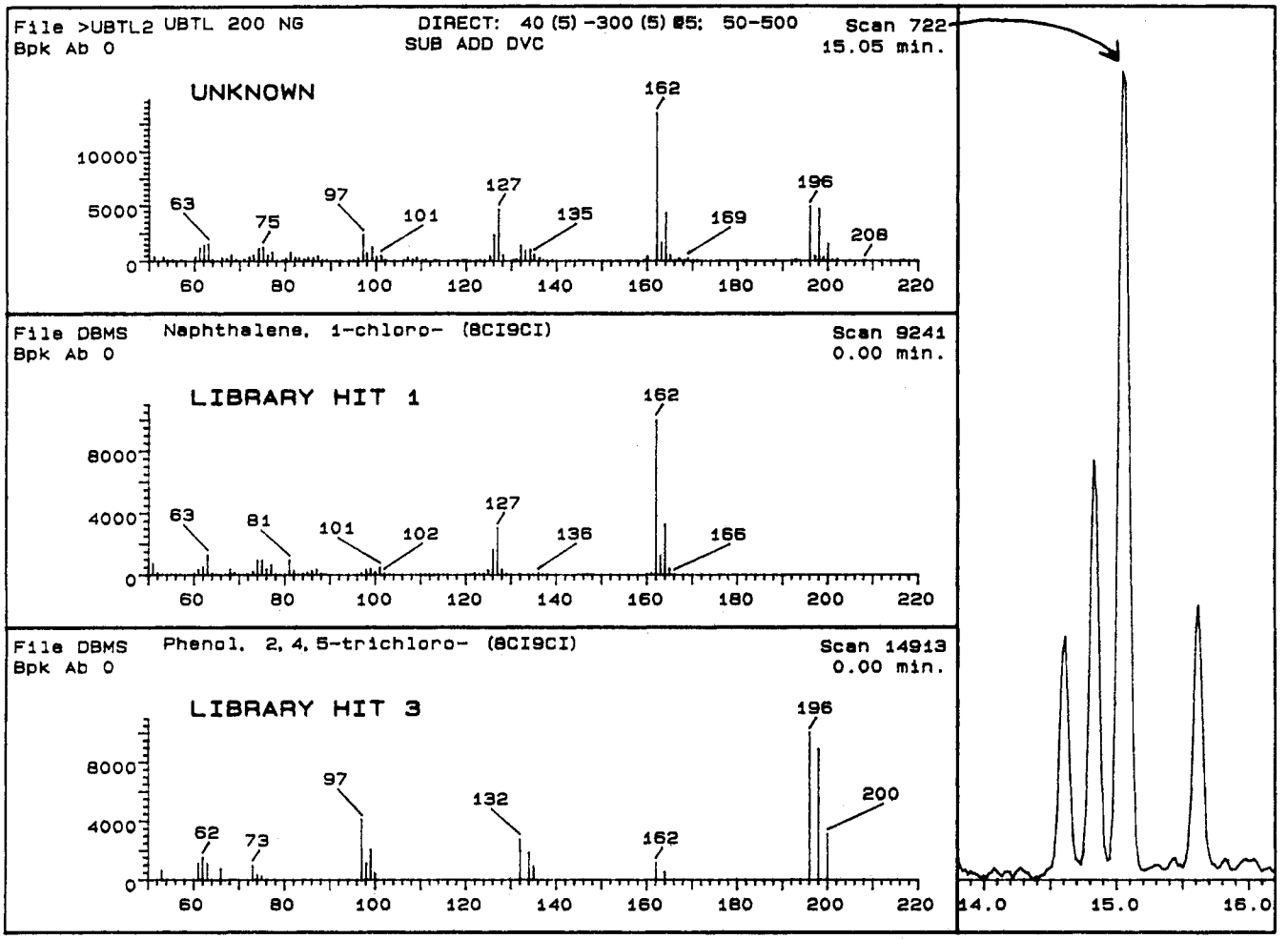


Figure 2. Retrieved reference spectra from real-time PBM matching of GC peak-top spectrum from Hewlett-Packard 5987 GC/MS system.

(Table IV); the uniqueness increase is obviously dependent on the mass accuracy of the measurement. A major problem is that essentially none of the current data bases of reference EI spectra was measured with the required mass accuracy. Thus, we are attempting to assign elemental compositions to these reference spectra through knowledge of the elemental composition of the molecule and of the most probable compositions of specific ions (e.g., mass 43) and neutrals lost. For example, important $(M - 15)^+$ and $(M - 36)^+$ ions from a reference of composition $C_8H_{10}Cl_2$ would most probably have the fragment ion compositions $C_7H_7Cl_2^+$ and $C_8H_9Cl^+$. For the latter, Table IV shows that only 13 of 400 *molecular ions* of the same mass have such a composition, a gain of 2^5 in the mass uniqueness of this peak (assuming that the weighting of m/z 140 fragment ions is the same as that for molecular ions).

GC/IR/MS. The structural information available from infrared absorption spectra is highly complementary to that from mass spectra. The exciting recent developments in Fourier-transform infrared instrumentation make it possible to obtain useful IR spectra on-line for 10^{-8} g of eluted GC components; practical GC/IR/MS systems have been demonstrated.[34] We are currently modifying PBM by adding vapor-phase infrared spectra, deriving weighting factors for frequencies and absorbances from a data base of 8000 reference spectra. In checking the above PBM answers wrong by class I but correct by class IV criteria (97%–84% = 13% for pure unknowns and 93%–79% = 14% for 60%-component unknowns), the majority appear clearly distinguishable on the basis of their infrared spectra.

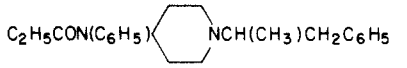
Matching Speed. Weighted ordering of the PBM reference file by the mass of the spectrum's most important peak (highest $U + A$ value) makes it possible to search only the most relevant reference spectra.[35] To avoid missing spectra of minor components of the unknown, reference spectra filed under masses of less important peaks (lower $U + A$) of the unknown are searched. Table V shows the effect of the $\Delta(U + A)$ of searching criteria on the percentage of the file searched, as well as the effect of filing references instead by the base peak mass. Obviously, weighting the file ordering by both mass and abundance is more effective than by abundance alone. A further speedup for low molecular weight compounds has been achieved by a secondary ordering according to the highest mass of a ≥3.4% abundance peak. This filing makes possible matching the 80 000 reference spectra for each component as it is eluted from the gas chromatograph for two or three components per minute. An example of such results from the Hewlett-Packard 5987 GC/MS system is shown in Figure 2. Secondary ordering based on the probability of a compound's occurrence (e.g., number of its spectra in the data base) should also give greatly increased speed if searching is terminated when a reasonably high (~85%) reliability match is achieved.

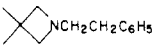
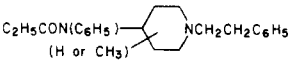
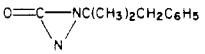
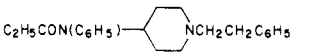
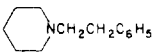
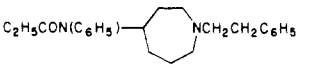
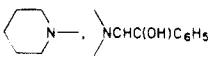
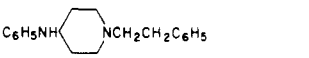
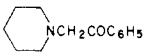
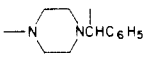
Table V. PBM with Weighted File Ordering

selection criteria	percent of file searched	x-fold speed increase	retrieval performance (%)
base peak[a]	13.5	7	100
$\Delta U + A = 3$[b]	7.5	13	100
$\Delta U + A = 2$[c]	3.8	25	99.2
$\Delta U + A = 1$	1.6	60	97.5

[a] Filing reference spectra under the mass value of the most abundant peak. [b] Searching spectra filed under the mass values of a minimum of 10 peaks in the unknown. [c] Correct answers retrieved under these conditions as a proportion of the 934 correct answers retrieved in searching the entire data base.

Table VI. Effect of Data-Base Size on STIRS Interpretation of China White Spectrum[a]



25 598 data base		61 023 data base	
best matching compound[b]	substructure	best matching compound[b]	substructure
1		1′, 3′, 5′	
2		2′	
3–5		4′	
6		6′	
7		7′	1
8		8′	2
		9′	3

[a] Best matching spectra using the "overall match factor", MF 11.0. Both data bases correctly predicted the molecular weight as 350. [b] The best matching compounds for the 25 598 data base were as follows: **1**, 3,3-dimethyl-1-phenethyl-2-phenylazetidine; **2**, 3-methoxy-6-methyl-17-(2-phenylethyl)-6α-diol; **4**, 1,2,3,4,5,6-hexahydro-6,11-dimethyl-3-(2-phenylethyl)-2,6-methano-3-benzazocin-8-ol; **5**, 3-methoxy-6-methyl-17-(2-phenylethyl)-6α-morphinan; **6**, 6-(6-acetonyl-5-hydroxy-1,4-dimethyl-5-phenyl-2-piperidyl)-3-hexen-2-one; **7**, 3-(4-hydroxy-4-methyl-1-piperidinyl)-1-phenyl-1-propanone; **8**, 1-[(4-chlorophenyl)phenylmethyl]-4-[[4-(1,1-dimethylethyl)phenyl]methyl]piperazine. For the 61 023 data base the best matching compounds were as follows: **1′**, 1-(2-phenethyl)-3-methyl-4-(*N*-propanilido)piperidine; **2′**, 2-methyl-1-phenethyl-4-(*N*-phenylpropionamido)piperidine; **3′**, 1-(2-phenethyl)-4-(*N*-propanilido)piperidine; **4′**, 1-phenethyl-4-(*N*-phenylpropionamido)azacycloheptane; **5′**, 1-(2-phenethyl)-2-methyl-4-(*N*-propanilido)piperidine; **6′**, *N*-phenyl-1-(2-phenylethyl)-4-piperidinamine.

SELF-TRAINING INTERPRETATIVE AND RETRIEVAL SYSTEM

Some interpretative systems, such as those utilizing "artificial intelligence",[3,5] attempt to teach the computer the necessary rules and spectral correlations to extract structural information from an unknown mass spectrum. "Pattern recognition" techniques have the computer itself derive and apply such correlations. The "K-nearest-neighbor" (KNN) approach[8] of the latter finds the position of the unknown spectrum relative to the references in a multidimensional mathematical hyperspace in which each peak mass is a dimension and its abundance is the distance in this dimension.

STIRS[6] can be viewed as a combination of these two approaches. Spectral rules and correlations have been used to define 26 data classes—combinations of masses or mass differences characteristic of different types of structural features. Then, for the unknown's data in each class the computer finds the most similar reference spectra (the nearest neighbors). If the proportion of best matching references containing a specific substructure is high relative to the proportion in the data base, this is indicative of the presence of that substructure in the unknown. For example, in Table VI for China White[36] the best matching of the 25 598 reference spectra contains a phenyl connected by two saturated carbons to a nitrogen in a saturated ring; the 61 023 data base gives additional best matching compounds that are much more definitive structurally. This STIRS "data weighting" based on known spectral correlations has been shown to give a substantially improved performance over that of the KNN method alone.[11] From an original list of >6000 substructures were selected the 589 best identified by STIRS for 900 unknowns chosen at random. Such statistical data also make it possible to assign the reliability of such substructural predictions by the 26 data classes.[37,38]

Other interpretive capabilities include identifying combinations of chlorine and bromine atoms present in peaks of the unknown by a PBM-type matching of their isotopic abundances.[39] Also the molecular weight of the unknown can be predicted (first choice, 91% accuracy; first or second, 94%) on the basis of the value yielding neutral-loss masses most closely resembling those of the STIRS-retrieval spectra.[40] The data base used by STIRS has been doubled to include reference spectra of 61 023 different compounds; this should substantially increase the proportion of substructures of the average unknown represented in these compounds. Table VI shows the best matching compounds found by STIRS for the unknown spectrum of the infamous China White illicit drug[36] before and after this doubling of the data base.

Future improvements planned that should increase the capabilities of STIRS in general are similar to those described for PBM. Weighted file ordering could give 1 order of magnitude improvement in search speed. Elemental composition data from exact mass measurement and, especially, infrared data should bring much better substructure prediction.

Hardware Improvements. The current Wiley/NBS data base requires ~24 megabytes of disc storage. However, even if the current data base update at Cornell should double its size, modern top of the line GC/MS systems should still have sufficient storage space. New hardware improvements such as the laser disc bring storage capacities far beyond any possible expansion of the MS and IR data bases.

Improved hardware could also greatly increase search speeds. For example, a parallel processor composed of several microprocessors could take full advantage of the inherently parallel nature of data-base searching. A processor composed of 10 of the newer "supermicro" processors should, in addition to providing instrument control capabilities, be able to do a PBM search in under 5 s while costing less than a laboratory minicomputer.

CONCLUSIONS

Instrument computerization has been one of the most remarkable revolutions in analytical capabilities in the last 2 decades. Use of the modern digital computer has progressed through data acquisition, data reduction, and feedback instrument control to data interpretation. The senior author worked with Roland S. Gohlke at the Dow Chemical Co. when he constructed and made operational in the middle 1950s what was probably the first GC/MS system. My memories of a

"successful" GC/MS run was the laboratory floor literally covered with UV-sensitive recording paper of dozens and dozens of mass spectra. The spectra could not be torn apart until numbering and labeling were completed. Then, masses had to be assigned and abundances measured before the human identification process could begin. This GC/MS system was highly impressive not only from its results but also from the fact that the results of 1 h could keep an interpreter and technician busy for days. For this problem, the development of on-line data acquisition and reduction systems was very helpful, but much more for the technician than the interpreter. Identification systems such as PBM and STIRS are now a real aid to the interpreter. The on-line GC/MS identification now produced by PBM greatly increases the interpreter's productivity, and its quantitative prediction of the reliability of the structure assignment can also alleviate the responsibility burdens of the interpreter. Confidence will be improved even further if reference spectra run under the same experimental conditions are available. The possibility that PBM will fail because the unknown spectrum is not represented in the data base is being reduced by increasing the data-base size. Further, for such cases the STIRS interpretive program can predict molecular weight, chlorine and bromine isotopic composition, and specific substructures present; for a list of 589 substructures, STIRS provides a quantitative prediction reliability.

The next stage of computer assistance has already been implemented for many specific problems. For example, for pollution, chemical exposure, and drug overdose problems the compound identification can be followed automatically by retrieval of its toxicological data.[41] For waste and dump-site screening, the identification could be followed by retrieval of the best disposal methods as well as precautions. For the medically revolutionary chemical diagnosis system of the future, compound identification could be followed by retrieval of the probable disease states. Such extensive automation of the analytical and interpretative process has the further key advantage of reducing unit costs, making much more available such critical help to these important societal problems.

ACKNOWLEDGMENT

In the development of the PBM and STIRS programs described here we are greatly indebted to Drs. R. Venkataraghavan, K.-S. Kwok, G. M. Pesyna, H. E. Dayringer, I. K. Mun, K. S. Haraki, B. L. Atwater (Fell), J. L. Serum, D. R. Bartholomew, W. Staedeli, R. D. Ellis, D. W. Peterson, M. Sharaf, R. B. Spencer, and S. Loh. The National Science Foundation, Grants 79-10400 and 83-03340, provided generous financial support.

REFERENCES AND NOTES

(1) Burlingame, A. L.; Whitney, J. O.; Russell, D. H. "Mass Spectrometry". *Anal. Chem.* **1984**, *56*, 417R.
(2) McLafferty, F. W.; Gohlke, R. S. "Mass Spectrometric Analysis. Spectral Data File Utilizing Machine Filing and Manual Searching". *Anal. Chem.* **1959**, *31*, 1160.
(3) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. "Applications of Artificial Intelligence for Chemical Inference. II. Interpretation of Low-Resolution Mass Spectra of Ketones". *J. Am. Chem. Soc.* **1969**, *91*, 2977.
(4) Hertz, H. S.; Hites, R. A.; Biemann, K. "Identification of Mass Spectra by Computer-Searching a File of Known Spectra". *Anal. Chem.* **1971**, *43*, 681.
(5) Smith, D. H.; Buchanan, B. G.; Engelmore, R. S.; Duffield, A. M.; Yeo, A.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. "Applications of Artificial Intelligence for Chemical Interference. VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids". *J. Am. Chem. Soc.* **1972**, *94*, 5962.
(6) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. "Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretative and Retrieval System". *J. Am. Chem. Soc.* **1973**, *95*, 4185.
(7) McLafferty, F. W.; Hertel, R. H.; Villwock, R. D. "Probability Based Matching of Mass Spectra. Rapid Identification of Specific Compounds in Mixtures". *Org. Mass Spectrom.* **1974**, *9*, 690.
(8) Justice, J. B.; Isenhour, T. L. "Information Content of Mass Spectra as Determined by Pattern Recognition Methods". *Anal. Chem.* **1974**, *46*, 223.
(9) Pesyna, G. M.; Venkataraghavan, R.; Dayringer, H. G.; McLafferty, F. W. "A Probability Based Matching System Using a Large Collection of Reference Mass Spectra". *Anal. Chem.* **1976**, *48*, 1362.
(10) Pesyna, G. M.; McLafferty, F. W. "Computerized Structure Retrieval and Interpretation of Mass Spectra". In "Determination of Organic Structures by Physical Methods"; Nachod, F. C.; Zuckerman, J. J.; Randall, E. W.; Eds.; Academic Press: New York, 1976; Vol. 6, pp 91–155.
(11) Lowry, S. R.; Isenhour, T. L.; Justice, J. B.; McLafferty, F. W.; Dayringer, H. E.; Venkataraghavan, R. "Comparison of Various K-Nearest Neighbor Voting Schemes with the Self-Training Interpretative and Retrieval System for Identifying Molecular Substructures from Mass Spectral Data". *Anal. Chem.* **1977**, *49*, 1720–1722.
(12) Henneberg, D. "Computerization and Library Search Systems". *Adv. Mass Spectrom.* **1980**, *8B*, 1511.
(13) Domokos, L.; Henneberg, D.; Wiemann, B. "Optimization of Search Algorithms for a Mass Spectra Library". *Anal. Chim. Acta* **1983**, *150*, 37.
(14) Cleij, P.; van't Klooster, H. A.; van Houwelingen, J. C. "Reproducibility as the Basis of a Similarity Index for Continuous Variables in Straightforward Library Search Methods". *Anal. Chim. Acta* **1983**, *150*, 23.
(15) Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. "Adding Forward Searching Capabilities to a Reverse Search Algorithm for Unknown Mass Spectra". *Anal. Chem.* **1985**, *57*, 771–773.
(16) Shackelford, W. M.; Kline, D. M.; Faas, L.; Kurth, G. "An Evaluation of Automated Spectrum Matching for Survey Identification of Waste Water Components by Gas Chromatography-Mass Spectrometry". *Anal. Chim. Acta* **1983**, *146*, 15.
(17) Cornell Computer Services, Uris Hall, Ithaca, NY 14853.
(18) Electronic Data Division, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158.
(19) McLafferty, F. W. "Performance Prediction and Evaluation of Systems for Computer Identification of Spectra". *Anal. Chem.* **1977**, *49*, 1441–1443.
(20) Salton, G. "Dynamic Information and Library Processing". Prentice-Hall: Ingelwood Cliffs, NJ, 1975.
(21) Atwater (Fell), B. L.; Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. "Reliability Ranking and Scaling Improvements to the Probability Based Matching System for Unknown Mass Spectra". *Anal. Chem.* **1985**, *57*, 899–903.
(22) Atwater (Fell), B. L.; Venkataraghavan, R.; McLafferty, F. W. "Mixture Spectrum Matching Utilizing Subtraction of Reference Spectra". *Anal. Chem.* **1979**, *51*, 1945–1949.
(23) Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. "Probability-Based Matching Algorithm with Forward Searching Capabilities for Matching Unknown Mass Spectra of Mixtures". *Anal. Chem.* **1985**, *57*, 1056–1060.
(24) McLafferty, F. W.; Stauffer, D. B. "An Improved Comprehensive Data Base for Matching Unknown Mass Spectra". *Int. J. Mass Spectrom. Ion Processes* **1984**, *58*, 139–149.
(25) Abramson, F. P. "Automated Identification of Mass Spectra by the Reverse Search". *Anal. Chem.* **1975**, *47*, 45.
(26) Baker, A. W.; Wright, N.; Opler, A. "Automatic Infrared Punched-Card Identification of Mixtures. Machine Method Combining Use of Band Wavelengths and Intervals of No Band". *Anal. Chem.* **1953**, *25*, 1457.
(27) Davis, J. M.; Giddings, J. C. "Statistical Theory of Component Overlap in Multicomponent Chromatograms". *Anal. Chem.* **1983**, *55*, 418.
(28) Page 118 of reference 20.
(29) Pesyna, G. M.; McLafferty, F. W.; Venkataraghavan, R.; Dayringer, H. E. "Statistical Occurrence of Mass and Abundance Values in Mass Spectra". *Anal. Chem.* **1975**, *47*, 1161–1164.
(30) Grotch, S. L. "Peak Height Distribution of Organic Mass Spectra". "17th Annual Conference", Dallas, May 1969; American Society of Mass Spectrometry; pp 459–466.
(31) Dromey, R. G. "Optimum Scaling of Mass Spectra for Computer-Matching". *Anal. Chim. Acta* **1979**, *112*, 133.
(32) Eichelberger, J. W.; Harris, L. E.; Budde, W. L. "Reference Compound to Calibrate Ion Abundance Measurements in Gas Chromatography-Mass Spectrometry Systems". *Anal. Chem.* **1975**, *47*, 995.
(33) Spencer, R. B.; McLafferty, F. W.; Stauffer, D. B.; Loh, S. "Improved Identification of Unknowns Using Enhanced Probability-Based Matching". "Annual Conference", New Orleans, February 1985; Pittsburgh Analytical and Spectroscopy Societies.
(34) Laude, D. A., Jr.; Brissey, G. M.; Ijames, C. F.; Brown, R. S.; Wilkins, C. L. "Linked Gas Chromatography/Fourier Transform Infrared/Fourier Transform Mass Spectrometry with Integrated Electron Impact and Chemical Ionization". *Anal. Chem.* **1984**, *56*, 1163.
(35) Mun, I. K.; Bartholomew, D. R.; Stauffer, D. B.; McLafferty, F. W. "Weighted File Ordering for Fast Matching of Mass Spectra against a Comprehensive Data Base". *Anal. Chem.* **1981**, *53*, 1938–1939.
(36) Cheng, M. T.; Kruppa, G. H.; McLafferty, F. W.; Copper, D. A. "Structural Information from Tandem Mass Spectrometry for China White and Related Fentanyl Derivatives". *Anal. Chem.* **1982**, *54*, 2204–2207.

(37) Haraki, K. S.; Venkataraghavan, R.; McLafferty, F. W. "Prediction of Substructures of Unknown Mass Spectra by the Self-Training Interpretive and Retrieval System". *Anal. Chem.* **1981**, *53*, 386-392.
(38) Sharaf, M.; Stauffer, D. B.; McLafferty, F. W., in preparation.
(39) Mun, I. K.; Venkataraghavan, R.; McLafferty, F. W. "Computer Assignment of Elemental Compositions of Mass Spectral Peaks from Isotopic Abundances". *Anal. Chem.* **1977**, *49*, 1723-1726.
(40) Mun, I. K.; Venkataraghavan, R.; McLafferty, F. W. "Computer Prediction of Molecular Weights from Mass Spectra". *Anal. Chem.* **1981**, *53*, 179-182.
(41) Heller, S. R. "Chemical Information System and Spectral Databases". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 224-231.

Structure Elucidation System Using Structural Information from Multisources: CHEMICS

SHIN-ICHI SASAKI*

Toyohashi University of Technology, Tempaku, Toyohashi 440, Japan

YOSHIHIRO KUDO

Faculty of Engineering, Yamagata University, Jonan, Yonezawa 992, Japan

Received February 1, 1985

One of the best ways of structure elucidation would be if a system could select the most probable structure from a gigantic file including all the possible structures that are known to exist, or that might exist from a chemical point of view, with the help of structural information, for example, a chemical spectrum. However, it is not possible to store all possible structures in a file. For example, even $C_{23}H_{48}$ has 5 731 580 isomeric structures. To overcome this, the present CHEMICS is designed to store all the substructures (called "components") necessary for building any likely structures. The set of components has been devised so that it is possible to construct any structure by selecting appropriate components from the component set. To store such a set of components in a computer is logically synonymous with storing all the complete structures that could be present. CHEMICS (CHEMICS-6) contains 189 components for the structure elucidation of organic compounds consisting of only C, H, and O atoms. A trial and error method was adopted in the selection of these components, with due regard to the prerequisites that the components should have no substructures overlapping one another and that the presence of the components could be deduced from structural information such as molecular formula and spectral data. Furthermore, 572 components have recently been prepared for CHEMICS-7 to handle samples that contain N, halogen, and S atoms in addition to C, H, and O. In the image space, all components are possible for an unknown compound (the analyst faces an almost infinite set of possible structures when he is without any structural information). CHEMICS-6 and -7 have the following tasks: (1) to eliminate unappropriate components, from the prepared component set, that are inconsistent with the molecular formula and spectral data; (2) to generate complete structural formulas from the retained components; (3) to exclude unlikely structures from those that have been generated; (4) to generate possible stereoisomeric structures, if a candidate structure possesses a stereocenter; (5) to output the most likely candidate structures, ideally the single correct solution. This paper describes the progress story up to the present CHEMICS. As to the details of the current CHEMICS, readers are requested to refer to the article that appeared in *Computer Enhanced Spectroscopy* (**1983**, *1*, 55).

Organic chemists acquainted with chemical information sciences have developed the ongoing system CHEMICS,[1-3] which is a total system of chemists, by chemists, and for chemists. The acronym stands for Combined Handling of Elucidation Methods for Interpretable Chemical Structures.[3] The system has been designed so as to enumerate exhaustively all possible structures consistent with given information for unknowns of moderate-sized structures. This system will work well to give more precise response by combination with a spectral file retrieval system (Appendix 1). CHEMICS has been improved step by step, and there are several versions of CHEMICS (the CHEMICS family) as shown in Figure 1. The value of each version of CHEMICS is not sequential but vectorial (Table I). The present paper will describe the main principles used in CHEMICS development.

Earlier versions of CHEMICS, as will be described later, presented only partial structures derived from spectral data input.[1,4,5] Now the system enumerates all possible structures by means of a set of possible components extracted from the list of whole components that are necessary and sufficient to build any kind of structure (Appendix 2, Figure 2).[6,7] The extraction is carried out by the comparison of the unknown's spectral data and molecular formula with the list followed by checking them by ^1H and ^{13}C NMR spectra again (Appendix 3).

AUTHORS' STAND ON THE SYSTEM CONTRACTION

Use of All Available Information from Multisources. CHEMICS makes it a general rule to utilize any relevant information that is available. However, use of much effective information is compelled to be postponed due to two main reasons. One is the limitation of the hardware available for our use. Because of this limitation, only a small part of the given information was utilized in CHEMICS. The limitation was particularly significant at the earlier stage of development, when only a computer with a cpu core of 4K 16-bit words was available. The other is the difference between styles of elucidation of human chemist and CHEMICS. Human chemists treat all things as working hypotheses because chemical knowledge is not always well-defined, so that a determined structure is only one of the possibilities afforded by a complex

0095-2338/85/1625-0252$01.50/0 © 1985 American Chemical Society