165. Applications of "Artificial Intelligence" for Chemical Inference, VI[1])

Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer

by **Armand Buchs[2]), Allan B. Delfino[3]). A. M. Duffield, Carl Djerassi, B. G. Buchanan, E. A. Feigenbaum** and **J. Lederberg**

Contribution from the Departments of Chemistry, Computer Science and Genetics, Stanford University, Stanford, California 94305, USA

(15. VI. 70)

Résumé. Le programme connu sous le nom de «Heuristic DENDRAL» est maintenant capable d'interpréter d'une manière absolument automatique les spectres de masse à basse résolution de n'importe quel composé de formule élémentaire $C_nH_{2n+v}X$ (X = 0, S ou N, v = valence de X). La possibilité de faire usage de spectres de RMN. pour faciliter l'interprétation a été retenue. Il n'est plus nécessaire de fournir au programme la formule élémentaire du composé dont on veut déterminer la structure. Les données théoriques concernant la spectrométrie de masse et la résonance magnétique nucléaire sont créées par le programme lui-même. A aucun moment le chimiste n'a besoin de fournir d'autres données que le spectre de masse et, s'il le désire, le spectre de RMN. L'efficacité du programme a été mise à l'épreuve avec 210 spectres de masse. La structure correcte apparaît toujours dans la réponse. Les résultats reportés dans les tableaux 2, 3 et 4 montrent que le nombre d'isomères qui sont compatibles avec la réponse donnée par le programme représente une très importante réduction du nombre total d'isomères qui sont *a priori* des candidats possibles.

Previous publications have described the results of heuristic computer programming for the interpretation of low resolution mass spectra of ethers [2] and amines [3]. These two classes of compounds are part of the general heteroatomic class $C_nH_{2n+v}X$

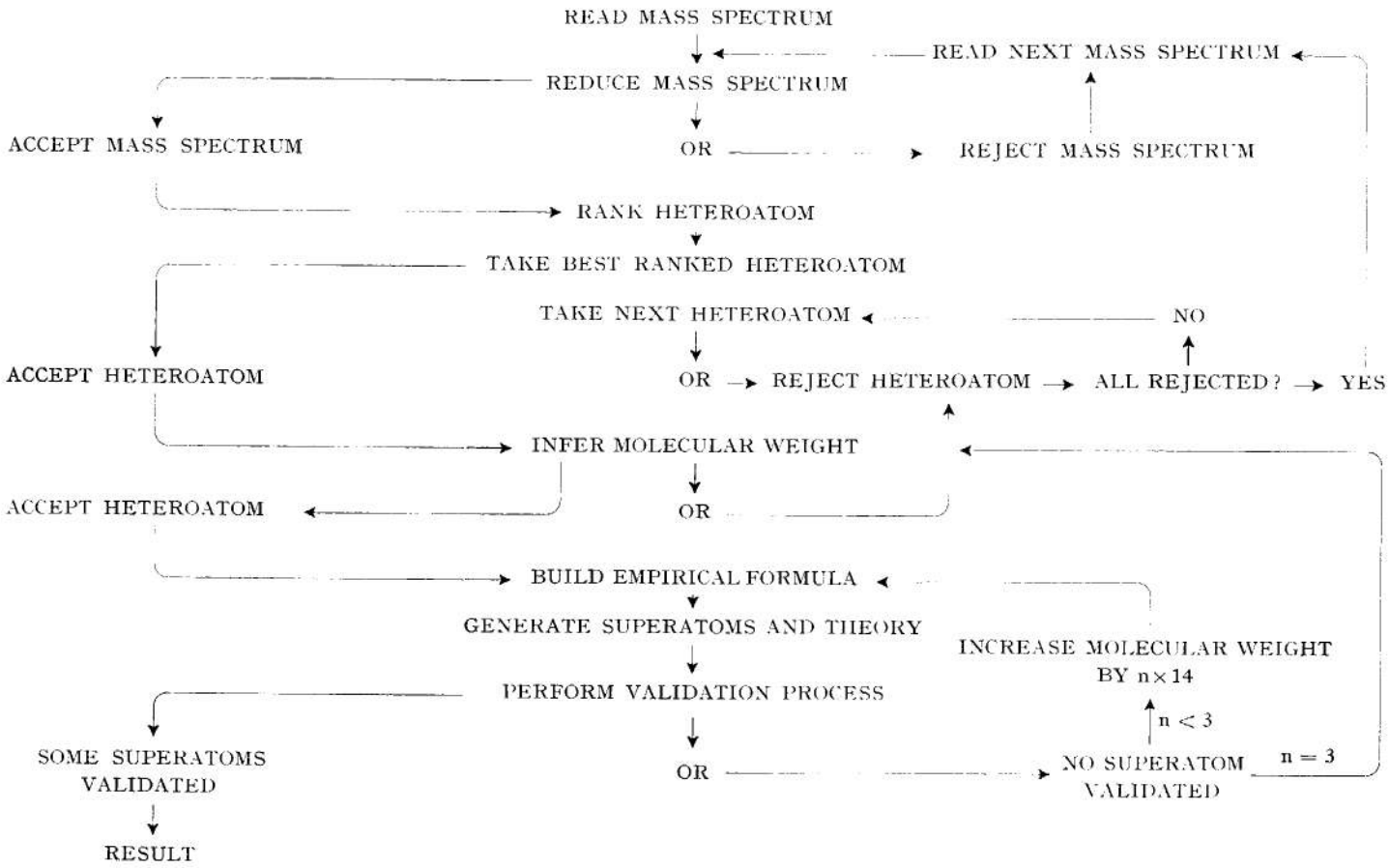
---

1) For Part V see reference [1].

2) On leave of absence from the University of Geneva, Switzerland.

3) Present address: *Allen-Babcock Computing*, Palo Alto, California 94303.

Diagram 1. *Choice of the most plausible empirical formula*



(v = valence of X) with which this paper is concerned. We shall review them in the light of improvements which have recently been achieved. The ether subclass that the program can analyze has been extended past methyl, ethyl and propyl ethers, to include *any ether structure*. Moreover, the alcohol, thioether, and thiol classes have been added to the program's repertoire. The necessity of supplying the empirical formula *has been removed*; the INFERENCE MAKER program is, at the present time, able to accept as sole inputs the mass spectrum and, optionally, the NMR. spectrum of the unknown compound. The purpose of this paper is to describe how the program first decides on a plausible empirical formula (and therefore a molecular weight), how it then generates the corresponding set of subgraphs, builds for each subgraph the theory related to its structure, and finally infers plausible substructures from the mass spectra of amines, ethers, alcohols, thiols, and thioethers.

The basic design of Heuristic DENDRAL is described in our earlier publication dealing with saturated ethers [2], and is summarized again in our publication dealing with amines [3]. As will be shown in this paper, the efficiency achieved in the INFERENCE MAKER with the general class of 'saturated acyclic monofunctional' (SAM) compounds is such that the two other phases of Heuristic DENDRAL (STRUCTURE GENERATOR and PREDICTOR) *need not to be used.*

Diagram 2. *INFERENCE MAKER output with heptane-3-ol* (**1**) *as an unknown*

ACTUAL MASS SPECTRUM = ((27.41) (28.11) (29.40) (30.3) (31.40) (32.1) (41.48) (42.6) (43.25) (44.6) (45.12) (55.13) (56.7) (57.18) (58.10) (59.100) (60.3) (67.1) (69.67) (70.5) (71.1) (72.1) (73.2) (84.1) (85.1) (86.2) (87.30) (88.2) (98.3))

MASS SPECTRUM CORRECTED FOR ^{13}C = ((27.41) (28.11) (29.40) (30.3) (31.40) (32.1) (41.48) (42.6) (43.25) (44.6) (45.12) (55.13) (56.7) (57.18) (58.10) (59.100) (60.1) (67.1) (69.67) (70.3) (71.1) (72.1) (73.2) (84.1) (85.1) (86.2) (87.30) (88.1) (98.3))

NMR. SPECTRUM = ((9.20 6T) (1.37 8M) (3.40 1M))

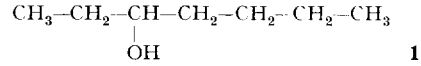
Run 1

NUMBER OF CARBON-BOUND METHYLS	= 2
NUMBER OF OXYGEN-BOUND METHYLS	= 0
TOTAL NUMBER OF METHYLS	= 2
MINIMUM NUMBER OF ALPHA-CARBON-BOUND HYDROGENS	= 1
INFERRED MOLECULAR WEIGHT	= 116
INFERRED EMPIRICAL FORMULA	= $C_7H_{16}O$
SUBGENERA INFERRED:	
*EA–S–(C_4H_9, C_2H_5)	1 ISOMER
TOTAL NUMBER OF ISOMERS:	*1*

Run 2

WAS A NMR. SPECTRUM AVAILABLE ?	NO
INFERRED MOLECULAR WEIGHT	= 116
INFERRED EMPIRICAL FORMULA	= $C_7H_{16}O$
SUBGENERA INFERRED:	
*EA–S–(C_4H_9, C_2H_5)	4 ISOMERS
TOTAL NUMBER OF ISOMERS:	*4*

The decision processes invoked by the INFERENCE MAKER in the choice of the most plausible empirical formula are schematically represented in Diagram 1, and will be illustrated with an example, the mass spectrum of heptane-3-ol (**1**).



1

The actual mass spectrum of **1** and the one corrected for the ^{13}C isotope contributions are tabulated in Diagram 2[4]). The program is supplied with the actual mass spectrum (and the NMR. spectrum, if one was recorded), and starts by making a decision about the plausibility of it belonging to the SAM class. The program strips from the mass spectrum all the ion signals which would be used later on, during the validation process. Then, depending on the average intensity[5]) of the remaining ion signals of the spectrum (called reduced spectrum), the program either accepts this mass spectrum as a plausible SAM candidate, or totally rejects it from further consideration at this very early stage of the process. For the SAM class compounds, the ions which are removed from the mass spectrum can all be formed by mechanistically important fragmentation paths. They belong to the following series:

1. α-Cleavage series for nitrogen, oxygen and sulfur SAM compounds starting with *m/e* 30 ($CH_2{=}\overset{+}{N}H_2$), 31 ($CH_2{=}\overset{+}{O}H$), and 47 ($CH_2{=}\overset{+}{S}H$) respectively. The following ions which belong to these series are removed from the mass spectrum of **1**: *m/e*: 30 (CH_4N)[6]), 31 (CH_3O), 44 (C_2H_6N), 45 (C_2H_5O), 58 (C_3H_8N), 59 (C_3H_7O), 72 ($C_4H_{10}N$), 73 (C_4H_9O), 86 ($C_5H_{12}N$), and 87 ($C_5H_{11}O$).

2. Alkyl series ions (C_nH_{2n+1}) arising from bond rupture between the heteroatom and an α-carbon with the charge remaining on the hydrocarbon moiety. This removes the ions with *m/e* 29 (C_2H_5), 43 (C_3H_7), 57 (C_4H_9), 71 (C_5H_{11}), and 85 (C_6H_{13}) from the actual mass spectrum of **1** (Diagram 2).

3. Alkyl series ions (C_nH_{2n-1}) originating from the primary loss of water and a methyl radical, followed by olefin expulsion. The ions with *m/e* 27 (C_2H_3), 41 (C_3H_5), 55 (C_4H_7), and 69 (C_5H_9) were also eliminated from the mass spectrum of **1**.

4. Alkyl series ions (C_nH_{2n}) arising from the loss of XH_2 (X = O or S), followed by expulsion of olefinic molecules. In the mass spectrum of **1** the following ions belong to that category: *m/e* 28 (C_2H_4), 42 (C_3H_6), 56 (C_4H_8), 70 (C_5H_{10}), and 98 (C_6H_{12}). They are therefore removed from the actual mass spectrum of **1**.

In order for a mass spectrum to be accepted as a plausible SAM candidate, the reduced spectrum must not only exhibit a low average intensity (< 3%), but must not contain any signal with an intensity greater than 10%. The reduced spectrum of **1** contains the following ions:

m/e:	60	67	88
Intensity:	3	1	2

[4]) The mass spectra reported in Diagr. 2 are tabulated in a sequence of dotted pairs. In each dotted pair the right part represents the relative abundance of the ion whose mass is given in the left part.

[5]) All intensity values refer to relative abundances with intensity of the base peak = 100%.

[6]) Since the program has not yet made a decision about the heteroatom, it considers the ions with *m/e* 30, 44, 58, 72, and 86 as arising by α-cleavage from an amine molecular ion; actually, in the mass spectrum of **1**, their empirical formulae are $C_nH_{2n}O$ (n = 1 to 4).

As it satisfies both these conditions, the actual mass spectrum tabulated in Diagram 2 is accepted as a SAM molecule spectrum and is subjected to further tests.

The program then assigns to each heteratom it knows, *i.e.* presently nitrogen, oxygen and sulfur, a plausibility score, by summing the intensities of the theoretical series of α-fission ions corresponding to each heteroatom. For any heteroatom X, the lowest mass α-cleavage peak has a mass corresponding to the formula $CH_2{=}XH_{v-1}$ (v = valence of X). In order to calculate the scores, the program uses the following mathematical relationships:

$$A = \text{Mass(X)} + \text{Valence(X)} + \text{Mass(CH)}$$

where X = Heteroatom,

$$M = A + (14 \times \text{i})$$

where i = one less than the carbon number corresponding to M_i.

J = Intensity of the ion of mass M,

$$\text{Score} = \sum_{\text{i}=0}^{\text{i}=\text{n}} J(M_i)$$

where n is defined by the following relation:

$$(14 \times \text{n}) + A < M_{max} < (14 \times (\text{n}+1)) + A$$

and M_{max} = Highest mass number present in the spectrum.

This score is calculated for each heteroatom. For the mass spectrum of **1** (see Diagram 2) the following scores are calculated from the above mentioned equations:

Nitrogen: Ions of m/e $(30+14 \times \text{i})$, i = 0 to 5: Sum = 22
Oxygen: Ions of m/e $(31+14 \times \text{i})$, i = 0 to 5: Sum = 184
Sulfur: Ions of m/e $(47+14 \times \text{i})$, i = 0 to 4: Sum = 0

The heteroatoms are then ranked in order of descending scores. With our example **1**, the program thus classifies oxygen as the most plausible, nitrogen as the next most plausible, and sulfur as the least plausible heteroatom.

Starting with the highest ranked heteroatom, the program then checks if its score exceeds a predefined minimum value[7]). If the score is lower than that minimum value, the spectrum under study cannot arise from a SAM class compound containing the highest ranked heteroatom in its structure, and the program proceeds to the next highest ranked heteroatom. This minimum value depends on the heteroatom. The so called α-cleavage ion series not only includes ions formed by α-cleavage, but also ions arising from cleavage occurring further away from the charge center, as well as ions formed by processes involving hydrogen migration. For example, in the spectrum of **1**, besides the two actual α-cleavage ions at m/e 59 (C_3H_7O) and 87 ($C_5H_{11}O$), the ion at m/e 45 (C_2H_5O) (which is formed by a rearrangement process[8])) and the one at m/e 73 (C_4H_9O) (which arises from β-cleavage) also belong to the so-called α-cleavage ion series used to rank the heteroatoms. All these ions contain the heteroatom in their structure. Hence, the better the heteroatom stabilizes the charge, the higher will be the sum of intensities of the ions found in that series. The score asso-

[7]) All threshold values were chosen on theoretical grounds. They were not optimized for the 210 mass spectra interpreted by the program, but adjusted so as to allow for a rather large safety margin.

[8]) See mechanism depicted under **2**, p. 1405.

ciated with nitrogen for example has to be greater than the predetermined value of 100% if nitrogen is to be kept for further tests. The minimum values for the scores associated with either oxygen or sulfur are respectively 5% and 20%; they need not be as high as for nitrogen, since these heteroatoms do not retain the positive charge as well as does nitrogen.

Once these preliminary tests have been performed, the program takes the best ranked heteroatom and makes a decision about the most plausible molecular weight. With our example the program starts with oxygen. Again, for any heteroatom, the molecular weight is to be found in the ion series given by the following relation:

Molecular weight = Mass(X) + Valence(X) + 14 × n

where n = Number of carbon atoms and X = Heteroatom.

Starting with the molecular weight of the lowest homolog ($CH_{2+v}X$), the program increases this value by steps of fourteen mass units until this value (M'_{max}) exceeds the mass of the last ion present in the ordered spectrum (M_{max}). It then either keeps this last value as the molecular weight or reduces it by fourteen mass units depending on the difference between M'_{max} and M_{max}. If this difference is larger than eleven mass units, the value of the lowest probable molecular weight is M'_{max} minus 14. Otherwise M'_{max} is taken as the lowest probable molecular weight. Moreover, if the inferred heteroatom is oxygen, the program checks if the value of ($M'_{max} - M_{max}$) equals 3 or 4. In such a case the program infers as lowest probable molecular weight the value ($M'_{max} + 14$). This takes into account the fact that, for many alcohol mass spectra, the last ion in the spectrum arises by the loss of water from the molecular ion. Evaluating the formula given above for this process, we find the following values from the mass spectrum of **1** with oxygen as the heteroatom:

Molecular weight = 16 + 2 + (14 × n)

$M_{max} = 98$

when n = 6, $M'_{max} = 102$, *i.e.* greater than M_{max}.

Since the value of ($M'_{max} - M_{max}$) equals 4, the program assumes that *m/e* 98 (C_7H_{14}) corresponds to the loss of H_2O and therefore adds 14 mass units to the value of M'_{max}, inferring *m/e* 116 as the lowest plausible molecular weight; it will use this value in order to eventually build the first empirical formula.

The results we have obtained with 210 mass spectra of amines, ethers, alcohols, thiols, and thioethers show that the correct molecular weight is always inferred on the first attempt for those mass spectra whose highest mass number is either M, $M-1$, $M-2$, $M+1$, $M+2$, $M+3$, or even $M-18$ and $M-17$ for oxygen containing compounds. The molecular ion *need not be present* in the spectrum. If the highest mass number in the spectrum is smaller than that of $M-10$, the program will infer a molecular weight M' of the next lower homolog, provided this does not lead to the apparent presence of intense ions at mass-spectrometrically improbable mass points $M'-R$ (with $2 < R < 15$). A mass spectrometrist would have to deal with this kind of spectrum in much the same manner as does the program. When the program is working with oxygen or with sulfur, it makes a final decision about allowing the spectrum to enter the validation process with one of these two heteroatoms. In the electron impact induced fragmentation of alcohols, ethers, thiols, and thioethers, the hydrocarbon moiety of the molecule plays an important role [4]. A rather large

fraction of the total ion current is carried by the hydrocarbon type ions C_nH_{2n+v} and C_nH_{2n-1}. To accept the spectrum with oxygen or sulfur as heteroatom, the program requires that the sum of the average intensities in the two above mentioned hydrocarbon series be greater than respectively 5% or 2%. The two ion series start with n = 3 [9]), *i.e.* with the ions *m/e* 41 and 43, and end when the value of n is such that *m/e* of ion C_nH_{2n+1} exceeds the mass of ion $(M - CH_2XH_{v-1})$. With our example (**1**) the C_nH_{2n-1} series includes the following ions: *m/e* 41, 55, 69, and 83. The average intensity value includes all the ions, *i.e.*, even those which are missing from the spectrum, such as *m/e* 83 with our example. The C_nH_{2n+1} series includes the ions of *m/e* 43, 57, 71, and 85. Since the sum of the average intensities of these two series amounts to 43%, *i.e.* to a value well above the 5% required, oxygen is accepted as a plausible heteroatom.

Once a molecular weight has been inferred, the program generates the empirical formula. Given the inferred heteroatom, the calculation is performed for SAM compounds in the following way:

if M = Inferred molecular weight, X = Heteroatom,
and C_nH_yX = General formula,
then $n = (M - \text{Mass(X)} - \text{Valence (X)})/14$
and $y = M - ((12 \times n) + \text{Mass(X)})$.

For example **1**, for which 116 was inferred as the value of M (with X = oxygen), this results in the following calculations:

$n = (116 - 16 - 2)/14 = 7$
$y = 116 - ((12 \times 7) + 16) = 16$
i.e. **Empirical formula = $C_7H_{16}O$**

After having built the empirical formula, the program builds the subgraphs or superatoms [10]) corresponding to that heteroatom and the theory associated with those superatoms. With our example **1**, the program generates the ether and alcohol subgraphs, formulating for each subgraph its associated mass and NMR. spectral theory, and tries to validate these subgraphs. If one or more subgraphs are validated, the total inference process for the unknown structure is complete; if no subgraph is validated for the molecular weight, the attempt is classed as a failure. Therefore the program makes a further attempt with the same heteroatom but a different molecular weight. Since the first molecular weight was a lower limit, the new molecular weight will be 14 mass units greater than the prior one. From this then is calculated a new empirical formula. A molecular weight or empirical formula change does not affect the number and kind of superatoms required for validation; the superatoms and theory are built *de novo* only if a heteroatom change occurs.

If, after having tried to validate subgraphs corresponding to the best ranked heteroatom with three consecutive empirical formulae, no substructure is substantiated, the program assumes that despite its high score, the highest ranked, and

[9]) The program ignores the two ions at *m/e* 27 and 29 (n = 2). In general they are of no value for the interpretation of mass spectra, especially with SAM compounds.

[10]) A superatom is defined as a structural unit with at least one free valence. In this context, the program generates only superatoms containing the heteroatom and all the α-carbon atoms with their protons; also, the program attaches only carbon atoms to the free valences.

accepted, heteroatom is not the correct one. The INFERENCE MAKER then makes the same kind of attempt with the next best ranked heteroatom, *i.e.* checks its consistency with the mass spectrum, infers a starting molecular weight in accord with the mass of the new heteroatom and the highest mass number of the mass spectrum, calculates an empirical formula, generates subgraphs and corresponding theory, and invokes the validation process. If no result is supported after all the heteroatoms that are known to the program have been postulated with three consecutive empirical formulae each, the mass spectrum cannot have resulted from a SAM compound, as far as the INFERENCE MAKER program is concerned.

In actual practice the program did find a subgraph consistent with the mass spectrum of heptane-3-ol (**1**). The actual output illustrated in Diagram 2 consists of two separate runs; in the first one the mass spectrum was supplemented by a NMR. spectrum, and in the second run the NMR. spectrum was ignored. If no subgraph had been validated for $C_7H_{16}O$, the program would have substituted $C_8H_{18}O$ and finally $C_9H_{20}O$. If still no subgraph were validated, the program would have classified the mass spectrum as not belonging to a compound of the SAM class. Nitrogen or sulfur subgraphs would not have been generated, because the observed scores (22 and 0) are below the threshold values (100 and 20) for both these two heteroatoms.

These preliminary decisions about consistency between heteroatom and spectrum *do not ensure* that only mass spectra of SAM compounds will enter the validation process, but they sharply decrease the probability of having non SAM compounds spectra accepted. If should be stressed that even if inadequate mass spectra pass that entrance filter, they still have to undergo successfully numerous tests during the validation process in order to be wrongly classified as SAM compound mass spectra.

Diagram 3. *Relations between the name and the structure of superatoms*

HETEROATOM PREFIXES		
EA = O	AM = N	TH = S
α-SUBSTITUTION SYMBOLS		
M = $-CH_3$	P = $-CH_2-$	S = $-\overset{\vert}{C}H-$	T = $-\overset{\vert}{\underset{\vert}{C}}-$
$-CH_2-O-CH_2-$	$-\overset{\vert}{C}H-O-CH_2-$	$HO-\overset{\vert}{\underset{\vert}{C}}-$
EA–PP	*EA–SP*	*EA–T*
$-CH_2-S-CH_3$	$-\overset{\vert}{C}H-S-CH_2-$	$-\overset{\vert}{\underset{\vert}{C}}-S-\overset{\vert}{\underset{\vert}{C}}-$
TH–PM	*TH–SP*	*TH–TT*
$\begin{matrix}\vert & \vert \\ CH-NH-CH \\ \vert & \vert\end{matrix}$	$-\overset{\vert}{C}-N-CH_3$ $\vert$ CH_3	$-\overset{\vert}{\underset{\vert}{C}}-N-CH-$ $\vert$ CH_2-
AM–SS	*AM–TMM*	*AM–TSP*

For each class of compounds, the subgraphs built by the program must represent a complete and irredundant set of substructures. Any SAM structure must belong *to one and only one* subgraph. This is accomplished by using for the superatom names a combination of the four symbols T, S, P and M called *α-substitution symbols* (see Diagram 3), preceded by a *heteroatom prefix* (AM for nitrogen, EA for oxygen, and TH for sulfur). The meaning of the α-substitution symbols and the structure each symbol or combination of symbols represents is described in our publication dealing with amines [3]. We will briefly review this notational scheme and illustrate it for the general class of SAM compounds.

For each subclass (amines, alcohols, ethers, thiols, thioethers) the number of superatoms depends on the valence of the heteroatom. For nitrogen, all combinations of the symbols T, S, P, and M, taken one at a time, two at a time, and three at a time, result in a total of 31 superatoms. Because oxygen and sulfur are divalent, there are only combinations of one and two letters for these heteroatoms. The canonical order of the α-substitution symbols (T > S > P > M) requires the higher value symbol to be written to the left of a lower value symbol; this allows only one way to write a particular name. A subgraph with one tertiary α-carbon, one secondary α-carbon, and one α-methyl radical should have its partial name written as TSM and not STM, MST, or MTS. The number of symbols in the name represents the number of carbon atoms directly bound to the heteroatom (α-carbons). With the heteroatoms which are currently known to the program, *i.e.* oxygen, sulfur, and nitrogen, names with 3 symbols can only represent superatoms of tertiary amines; those with 2 symbols refer to secondary amines as well as to ethers and thioethers, while one-symbol names may represent primary amines, alcohols, and thiols. In each particular name, the α-substitution symbols themselves give the number of β-carbon(s) attached to each α-carbon atom (3 for T, 2 for S, 1 for P, and none for M). The general relationship between superatom names and the structure they represent is depicted in Diagram 3, along with some examples.

Once the INFERENCE MAKER has inferred a heteroatom, it builds the corresponding superatom names and for each superatom the program constructs a set of properties associated with the superatom and the mass spectrometric and NMR. related conditions which will have to be satisfied in order for the superatom to be validated. This is possible because the name of a superatom represents all the needed information (structure, weight, mass of the lowest possible α-fission peak, etc.). Moreover, the name of a superatom contains enough structural information to decide what kind of fragmentation can be expected to occur predominantly from a molecular ion containing as a subunit the partial structure represented by the name of that superatom.

The program builds a set of numbers using the digits 1, 2, 3, and 4. These numbers are allowed to contain from one to **n** digits, **n** being the valence of the heteroatom. No digit of a higher value can be written to the right of a digit of a lower value; all possible numbers that do not violate the canonical order must be included in the set. With example **1** the following 14 numbers are generated[11]: 1, 2, 3, 4, 11, 21, 22, 31,

[11] For n = 3 (*e.g.* nitrogen), the following 20 combinations would be added to the 14 generated for divalent heteroatoms: 111, 211, 221, 222, 311, 321, 322, 331, 332, 333, 411, 421, 422, 431, 432, 433, 441, 442, 443, and 444.

32, 33, 41, 42, 43 and 44. Each number is then translated to its corresponding α-substitution symbol (1 to M, 2 to P, 3 to S, 4 to T). The heteroatom prefix is attached with an intervening dash and the name is surrounded by two asterisks. The result is a name like *AM–SS* for the secondary amine superatom with both α-carbons mono-substituted (see Diagram 3). Since we are interested in subgraphs with at least one free valence, names containing only M's are ignored[12]).

Each superatom has intrinsic properties as well as properties connected with mass and NMR. spectrometry. Some of the properties depend only on the heteroatom prefix; they are constants for a given heteroatom. Some of the intensity threshold values used during the validation process are examples of such properties. Other properties depend only on the combination of the α-substitution symbols; they are not related to any particular heteroatom. Finally, each superatom has properties which are implied by both the heteroatom prefix and the combination of the α-substitution symbols. Moreover, if some properties simply are numerical values which the program will use to perform calculations, others represent switches which will tell the program what kind of tests to perform for each particular superatom.

The properties associated with each superatom are calculated and classified according the following outline:

A. *Intrinsic properties.* Structure and weight are the only two intrinsic properties; their value depends on the complete superatom name (heteroatom prefix and α-substitution symbol). The program knows the partial structure corresponding to each α-substitution symbol (M = $-CH_3$, P = $-CH_2-$, S = $-\overset{|}{C}H-$ and T = $-\overset{|}{C}-$). The heteroatom is deduced from the heteroatom prefix (AM = N, EA = O and TH = S) and the number of hydrogen atoms attached to the heteroatom is equal to the difference between the number of α-substitution symbols and the valence of the superatom. The weight of the superatom is not calculated from the chemical structure, but directly from the name. A mass is assigned to each α-substitution symbol (15 to M, 14 to P, 13 to S and 12 to T) and also to each heteroatom prefix. The mass corresponding to the various heteroatom prefixes is given by the mass of the molecule XH_v (v = valence of X). This results in the following values: 17 for AM, 18 for EA and 34 for TH. The mass of any superatom is obtained by adding the masses of the α-substitution symbols to the difference between the weight of the heteroatom prefix and the number of α-substitution symbols. For superatom *TH–SP* for example (see Diagram 3), this leads to the following calculation: $13+14+(34-2) = 59$.

B. *Mass spectrometric properties which depend on the α-substitution symbols only.* The number of carbon-carbon bonds available for α-cleavage or, equivalently, the number of free valences of the superatoms, and the total substitution degree of the α-carbons are examples of such properties. In order to calculate the number of free valences, the program assigns to each α-substitution symbol a value (0 to M, 1 to P, 2 to S and 3 to T). The sum of the values of each α-substitution symbol represents the number of free valences. For example, superatom *AM–TSP* (see Diagram 3) has $(3+2+1)$ *i.e.* 6 free valences.

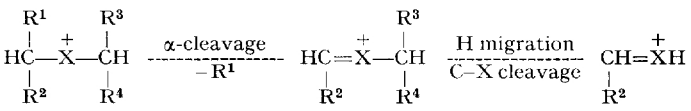
[12]) The three general names *X–M*, *X–MM* and *X–MMM* with X = AM, or EA and TH when at maximum two M's are present, represent molecules. *EA–M* and *EA–MM*, for example, stand for methanol and dimethyl ether respectively.

Table 1. *Tests used during the validation process*

SUPERATOMS	SIZE	NMR. tests			Mass spectrometry tests									
					Direct tests						Multistep tests			
		TMC	HMC	HYC	$M-XH_2$			$M-CH_3XH$	$CH_2=XH$	$CH_2=XCH_2$	EVION	ALPHA	REARR	ALKFIT
					<1%	>2%	1–100	>2%	>10%	>10%				
X–P	2	1	0	2	–	+	–	–	+	–	+	–	–	–
X–S	3	2	0	1	–	–	+	–	–	–	–	+	–	+
X–T	4	3	0	0	–	–	–	–	–	–	–	+	–	+
X–PM	3	2	1	2	+	–	–	+	–	+	–	–	–	+
X–PP	4	2	0	4	+	–	–	–	–	–	–	+	–	+
X–SM	4	3	1	1	+	–	–	–	–	–	–	+	–	+
X–SP	5	3	0	3	+	–	–	–	–	–	–	+	+	+
X–SS	6	4	0	2	+	–	–	–	–	–	–	+	+	+
X–TM	5	4	1	0	+	–	–	–	–	–	–	+	–	+
X–TP	6	4	0	2	+	–	–	–	–	–	–	+	+	+
X–TS	7	5	0	1	+	–	–	–	–	–	–	+	+	+
X–TT	8	6	0	0	+	–	–	–	–	–	–	+	+	+

X = EA or TH.
\+ means that the switch for that test is 'on'.
– means that the switch for that test is 'off'.

The total degree of substitution of the α-carbons represents a different kind of property. It constitutes a switch that the program sets 'on' or 'off', depending on the name of the superatom under test. During the validation process the program will perform some tests related to that property only if the switch is 'on'. In Table 1 are reported all the switches used for the validation of the 12 oxygen or sulfur superatoms. From now on they will be referred to as tests rather than switches. Some tests are simple ones, like checking the intensity of a particular ion signal (test '$M-XH_2$', Table 1), while others imply more complex multistep processes, like searching the mass spectrum for sets of α-cleavage ions at *m/e* consistent with the structure of the superatom under test, and having intensities in accord with the charge retentive power of the heteroatom (test 'ALPHA', Table 1). More extensive comment on test ALPHA will be made later in the text. The test 'REARR' for example (see Table 1), is set to the position 'on' for those superatoms which, if they were present in the molecular ion as the central subunit, would lead after electron impact to a favored hydrogen rearrangement process. This occurs only with molecular ions containing as part of their structure a superatom with at least one substituted α-carbon. For such molecular ions one can expect the mass spectrum to exhibit strong signals for ions arising from the well known [5] rearrangement mechanism depicted below (**2**, **b** → **c**) with an ether (X = O) or thioether (X = S) molecular ion as example:



a **b** **c**

2

Only superatoms with names containing at least two α-substitution symbols (excluding M's), with at least one of them being S or T, possess the required structure. To decide for which superatom the test should be performed, the program removes the M's from the superatom name and sets test REARR to 'on' or 'off' depending on which α-substitution symbols are left.

C. *Mass spectrometric properties which depend on the complete superatom name.* Examples of such properties include both tests and numerical properties. The lowest possible mass of an ion formed by α-fission for a particular superatom is an example of a numerical property. The program calculates the value of this property, for each superatom, by adding to the mass of the superatom the mass corresponding to (**n** −1) methyl radicals, where **n** represents the number of free valences. For superatom *TH–TT* for example (see Diagram 3), the smallest α-fission fragment is $(CH_3)_3$–C–S–C–$(CH_3)_2$–; it cannot have a mass smaller than *m/e* 131 (mass of superatom = 56, **n** = 6). An example of a test is represented by 'ALPHA' (see Table 1); it tells the program how to handle conditions related to α-cleavage, depending on the charge retentive power of the heteroatom and the structure of the superatom. The subtests it implies are described in the part dealing with the validation phase of the INFERENCE MAKER program. Other tests are simple intensity checks (tests 'CH_2=XH', 'CH_2=XCH_3', '$M-CH_3XH$', etc., Table 1).

D. *Mass spectrometric properties which only depend on the heteroatom prefix.* These properties include some of the various threshold values assigned to the intensity of particular ions or ion series. Oxygen containing superatoms, for example, are accepted for further consideration only if the hydrocarbon type ions C_nH_{2n-1} originating from C–O cleavage exhibit a sum of intensities greater than 5%. The program sets this threshold to different values for sulfur or nitrogen containing superatoms.

E. *Properties pertaining to NMR. spectrometry.* Here again, the values assigned to some of these properties depend only on the α-substitution symbols, while for others they change from heteroatom to heteroatom. Properties which have different values for different structures around the heteroatom are:

1. The minimum number of methyl radicals required by the structure of a superatom (test 'TMC', Table 1).

2. The number of methyl radicals linked to the heteroatom (test 'HMC', Table 1).

3. The maximum number of protons bound to α-carbon atoms, excluding methyl protons (test 'HYC', Table 1).

Since we are dealing exclusively with saturated chemical structures, the minimum number of methyl radicals that an NMR. spectrum should exhibit to congrue with a superatom structure is equivalent to the number of free valences added to the number of M's present in the name of the superatom. For example, the structure of superatom *AM–TMM* (see Diagram 3) requires that at least five methyl groups be inferred from the NMR. spectrum[13]). To calculate the number of methyl groups compatible with the structure of a superatom, the program simply counts the M's appearing in the name. A definite number of protons is part of the structure of every α-substitution symbol which has at least one free valence left for a carbon-carbon linkage (2 for P, 1 for S and 0 for T). By adding together all the protons of these α-substitution symbols, the program determines the maximum number of α-carbon hydrogens allowed by each structure. Superatom *EA–PP* for example (see Diagram 3), is assigned four such protons.

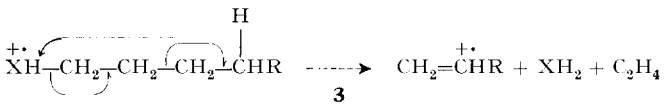
Once the superatom and theory generation phase has been completed, the program corrects the relative abundances of the signals in the mass spectrum by removing isotope peaks; it then deletes from the spectrum any peak appearing at an improbable mass ($M-3$ through $M-14$), adjusts the intensities of the remaining ions with respect to 100% for the base peak, and initiates the validation process for each of the 31 (nitrogen)[14]) or 12 (oxygen and sulfur) superatoms.

With oxygen or sulfur SAM compounds some of the tests are similar to those which were designed for amines; this holds for all the tests that are not related to mass spectrometry. The main difference arises from the fact that nitrogen, in contrast to either oxygen or sulfur, is very efficient in stabilizing, and hence retaining, the positive charge. This affects drastically the fragmentation pattern for amines, and as is shown in our publication [3], almost all the tests dealing with mass spectro-

[13]) In order to generate a SAM molecule from *AM–TMM*, the addition of three alkyl radicals is required. They could be methyl radicals or not, but, in this latter case, each alkyl chain must terminate in at least one methyl group.

[14]) A detailed description of the tests each amine superatom undergoes is given in our publication [3].

metry relied on the charge localization concept [6]; α-cleavage and rearrangement according to the mechanism previously depicted (see **2**) were the two main processes used by the INFERENCE MAKER program to efficiently interpret amine low resolution mass spectra. As is well known [7], oxygen and sulfur are less effective than nitrogen in accomodating the positive charge. α-Cleavage plays a less important role, especially when the size of the molecule, or the branching of the alkyl radicals, is substantial. The influence of the heteroatom upon the fragmentation is often overshadowed by the hydrocarbon moiety of the molecule; this has to be overcome for a successful interpretation of the mass spectrum. The partial lack of charge retention apparently hinders the ease of interpretation more for ethers and thioethers than for alcohols or mercaptans. The fragmentation is no longer triggered by a clear driving force as it was for amines. Other fragmentation paths have to be considered, like C–X bond scissions with the charge remaining on the alkyl radical (X = O or S), loss of XH_2, or HXR, followed by olefin expulsion according to the mechanism depicted below (**3**).



3

In order to describe how the validation phase of the INFERENCE MAKER program infers the correct superatom along with the size of the alkyl radicals C_nH_{2n+1} attached to each free valence, the various tests reported in Table 1 will be illustrated by using the mass spectrum of isopropyl *n*-amyl ether (**4**), a molecule which contains an *EA–SP* subgraph (see Diagram 3).

$$(CH_3)_2{-}CH{-}O{-}CH_2{-}(CH_2)_3{-}CH_3$$

4

The correct answer for that compound is: *EA–SP–(CH_3, CH_3) (C_4H_9), where EA stands for oxygen and SP gives the number and the structure of the α-carbon atoms.

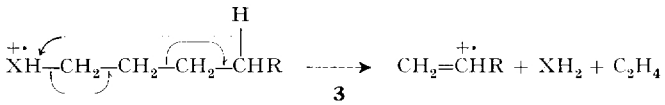
Diagram 4. *INFERENCE MAKER output with isopropyl* n-*amyl ether* (**4**) *as an unknown*

ACTUAL MASS SPECTRUM = ((31.2) (41.15) (42.10) (43.100) (44.4) (45.30) (55.6) (56.1) (57.1) (59.3) (69.3) (70.5) (71.43) (72.2) (73.21) (115.16) (116.1))

MASS SPECTRUM CORRECTED FOR ^{13}C = ((31.2) (41.15) (42.10) (43.100) (44.2) (45.30) (55.6) (56.1) (57.1) (59.3) (69.3) (70.5) (71.43) (72.1) (73.21) (115.16))

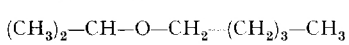
WAS A NMR. SPECTRUM AVAILABLE ?	NO
INFERRED MOLECULAR WEIGHT	= 116
INFERRED EMPIRICAL FORMULA	= $C_7H_{16}O$
SUBGENERA INFERRED:	*NONE*
WAS A NMR. SPECTRUM AVAILABLE ?	NO
INFERRED MOLECULAR WEIGHT	= 130
INFERRED EMPIRICAL FORMULA	= $C_8H_{18}O$
SUBGENERA INFERRED:	
*EA–SP–(CH_3, CH_3) (C_4H_9)	4 ISOMERS
TOTAL NUMBER OF ISOMERS:	*4*

metry relied on the charge localization concept [6]; α-cleavage and rearrangement according to the mechanism previously depicted (see **2**) were the two main processes used by the INFERENCE MAKER program to efficiently interpret amine low resolution mass spectra. As is well known [7], oxygen and sulfur are less effective than nitrogen in accomodating the positive charge. α-Cleavage plays a less important role, especially when the size of the molecule, or the branching of the alkyl radicals, is substantial. The influence of the heteroatom upon the fragmentation is often overshadowed by the hydrocarbon moiety of the molecule; this has to be overcome for a successful interpretation of the mass spectrum. The partial lack of charge retention apparently hinders the ease of interpretation more for ethers and thioethers than for alcohols or mercaptans. The fragmentation is no longer triggered by a clear driving force as it was for amines. Other fragmentation paths have to be considered, like C–X bond scissions with the charge remaining on the alkyl radical (X = O or S), loss of XH_2, or HXR, followed by olefin expulsion according to the mechanism depicted below (**3**).



3

In order to describe how the validation phase of the INFERENCE MAKER program infers the correct superatom along with the size of the alkyl radicals C_nH_{2n+1} attached to each free valence, the various tests reported in Table 1 will be illustrated by using the mass spectrum of isopropyl *n*-amyl ether (**4**), a molecule which contains an *EA–SP* subgraph (see Diagram 3).



4

The correct answer for that compound is: *EA–SP–(CH_3, CH_3) (C_4H_9), where EA stands for oxygen and SP gives the number and the structure of the α-carbon atoms.

Diagram 4. *INFERENCE MAKER output with isopropyl* n*-amyl ether* (**4**) *as an unknown*

ACTUAL MASS SPECTRUM = ((31.2) (41.15) (42.10) (43.100) (44.4) (45.30) (55.6) (56.1) (57.1) (59.3) (69.3) (70.5) (71.43) (72.2) (73.21) (115.16) (116.1))

MASS SPECTRUM CORRECTED FOR ^{13}C = ((31.2) (41.15) (42.10) (43.100) (44.2) (45.30) (55.6) (56.1) (57.1) (59.3) (69.3) (70.5) (71.43) (72.1) (73.21) (115.16))

WAS A NMR. SPECTRUM AVAILABLE ?	NO
INFERRED MOLECULAR WEIGHT	= 116
INFERRED EMPIRICAL FORMULA	= $C_7H_{16}O$
SUBGENERA INFERRED:	*NONE*
WAS A NMR. SPECTRUM AVAILABLE ?	NO
INFERRED MOLECULAR WEIGHT	= 130
INFERRED EMPIRICAL FORMULA	= $C_8H_{18}O$
SUBGENERA INFERRED:	
*EA–SP–(CH_3, CH_3) (C_4H_9)	4 ISOMERS
TOTAL NUMBER OF ISOMERS:	*4*

The second part of the answer indicates that two methyl radicals are attached to the 'S' α-carbon atom and a butyl radical (1-butyl, *sec*-butyl, *t*-butyl or isobutyl) to the 'P' α-carbon atom.

Each of the 12 oxygen superatoms built by the program is initially put on a list. The program then checks each superatom for consistency with the data (mass spectrum and NMR. spectrum if one was supplied). As soon as a superatom fails to pass a test, it is removed from the list. The final result shows all the remaining superatoms and, for each of them, the alkyl radicals attached to each free valence. Diagram 4 contains the mass spectrum of **4** and the answer given by the INFERENCE MAKER on the basis of that spectrum.

The first test (test 'SIZE', Table 1) is related to the size of the empirical formula which the program deduced from the mass spectrum. To pass that test, a superatom must not require more carbon atoms than are available. The minimum number of carbon atoms required by the structure of each superatom in order to build the smallest possible molecule is calculated by the program by adding the number of free valences to the number of α-substitution symbols; these minimum numbers are reported in Table 1 for each superatom. For $C_nH_{2n+2}X$ compounds (X = O or S), all superatoms pass that test provided n is greater than 7. With our example (**4**), the program selected $C_8H_{18}O$ as the second empirical formula, and no pruning was achieved by that test. For heptane-3-ol (**1**), superatom *EA–TT* is eliminated at that very early stage of the validation process.

The next three tests are only effective when an NMR. spectrum is supplied, in which case they are employed prior to any mass spectrometry tests. In order to build a saturated molecule, each superatom requires a minimum number of methyl radicals (test 'TMC', Table 1), a definite number of methyl radicals linked to the heteroatom (test 'HMC', Table 1), and a maximum number of α-carbon bound hydrogen atoms (test 'HYC', Table 1). Any superatom for which one of these conditions is not satisfied by the signals present in the NMR. spectrum is discarded from further consideration and will henceforth not be tested against the mass spectral data. It should be stressed that the program uses NMR. spectra *only as methyl counters* and, if desired, as α-carbon proton counters[15]. It does not rely on fully interpreted NMR. spectra; if the user has some doubts about the multiplicity of signals, or if no integration curve was recorded, the program will also accept partial information [3].

From the NMR. spectrum of heptane-3-ol (**1**) the program inferred the presence of two carbon-bound methyl radicals and no oxygen-bound methyl group (see Diagram 2, run 1). Superatoms *EA–PM*, *EA–SM* and *EA–TM*, which require the presence of a methoxy group, as well as all superatoms for which more than two methyl radicals are mandatory (see test 'TMC', Table 1) are eliminated by the NMR. filter. Only superatoms *EA–P*, *EA–S* and *EA–PP* pass. With that particular compound, the same final result is obtained with and without the aid of NMR. data, as far as the number of inferred superatoms is concerned (see Diagram 2). Using NMR. data results in an efficient pruning at the very beginning of the validation phase, and assigns a straight chain structure to the C_4H_9 radical. As no NMR. spectrum is recorded for isopropyl *n*-amyl ether (**4**), the program simply skips the NMR. tests.

[15] A detailed description explaining how the program takes advantage of NMR. data is reported in our previous publication dealing with amines [3].

The program then encounters the mass spectrometry tests[16]). The first condition programmed in the mass spectrometry part of the validation process is depicted in Table 1 as '$M-XH_2$'. If the peak at *m/e* corresponding to the mass of the $M-XH_2$ ion appears with an intensity greater than 1%, all superatoms with names formed by more than one α-substitution symbol are rejected. Mass spectra of secondary alcohols are allowed to display intensities between 1% and 100% for the $M-H_2O$ ion, and those of tertiary alcohols any intensity (from 0% to 100%) for that ion, but for primary alcohols this ion must be present in the spectrum with a relative abundance greater than 2%. For superatom *X–P* (X = EA or TH), the program then requires that the only peak which can arise from α-cleavage exhibits an intensity above 10% (test 'CH_2=XH', Table 1); if it does, the program calculates the average intensity of all ions belonging to the series $((M-XH_2)-C_2H_4 \times n)$, starting with n = 1 and ending at *m/e* 42 (test 'EVION', Table 1). If the average intensity exceeds 10% (20% for mercaptans), the program then checks the average intensity of ions C_nH_{2n+1} and C_nH_{2n-1}, starting with n = 3 (*m/e* 41 and 43) and increasing n until *m/e* of ion C_nH_{2n-1} equals the mass of $M-CH_2XH$, where *M* represents the molecular weight. Superatom *X–P* is definitely accepted if this last value exceeds 50% when X = EA or 85% when X = TH. The mass spectrum of **4** does not exhibit an $M-18$ ion. Superatoms *EA–P* and *EA–S* are therefore eliminated. Methyl ethers with a mono-substituted α-carbon always expell CH_3OH (32 mass units) upon electron impact; superatom *EA–PM* is rejected because no $M-32$ ion appears in the mass spectrum of **4** (test '$M-CH_3XH$', Table 1).

The next tests programmed into the validation process pertain to conditions about α-cleavage ions and the corresponding C_nH_{2n+1} ions formed by fission of the C–X bond. For those superatoms which have only one free valence, the program requires an intensity greater than 10% for the only possible α-cleavage ion (test 'CH_2–XH' and 'CH_2=XCH_3', Table 1). For any other superatom the program then builds all genera[17]) in accord with the structure of the superatom and the empirical formula. In order to achieve that, the masses of all theoretically possible α-cleavage peaks are calculated. If **n** represents the number of free valences of a superatom, *m/e* of the lowest mass α-fission ion which can be pictured by using the superatom's structure and the elements of the empirical formula is given by adding the mass of the superatom (*m*) to the mass of n–1 methyl radicals; *m/e* of the heaviest potential α-fission ion corresponds to the mass of the $M-15$ ion (*M* = inferred molecular weight). Considering superatom *EA–SP* (**n** = 3, *m* = 43), and empirical formula $C_8H_{18}O$, potential α-scission ions can only have the following masses: *m/e* = 73 (C_4H_9O), 87 ($C_5H_{11}O$),

[16]) Only tests for oxygen or sulfur SAM compounds will be discussed here. Those pertaining to amines have been extensively explained in our publication [3] and are still valid.

[17]) A generic description or *genus* is defined as an entity displaying the superatom and the alkyl radicals available for saturating the free valences, without any specification about the precise distribution of these radicals among the free valences. For example, *EA–SP(CH_3,CH_3,C_4H_9) is referred to as a genus. A description in which the respective positions of the radicals are unequivocally specified will be referred to as a *subgenus*. From the genus *EA–SP–(CH_3,CH_3,C_4H_9), the two subgenera *EA–SP–(CH_3,CH_3)(C_4H_9) and *EA–SP–(CH_3,C_4H_9)(CH_3) can be formed. Subgenera represent structures which are completely defined, with the exception of the inner structure of the C_nH_{2n+1} radicals attached to the α-carbon atoms when these radicals contain more than two carbon atoms.

101 ($C_6H_{13}O$) and 115 ($C_7H_{15}O$). From these masses, the program then calculates all combinations of **n** peaks which satisfy the following mathematical relationships:

If **M** = Molecular weight, **m** = mass of the superatom and $\mathbf{p}_i$ = *m/e* of an α-cleavage peak, then ($\mathbf{p}_i$, $\mathbf{p}_{i-1}$, ..., $\mathbf{p}_n$) with $\mathbf{p}_i < \mathbf{p}_{i+1} < \ldots < \mathbf{p}_n$ is a valid combination if the equation $\boxed{\sum_{i=1}^{i=n} \mathbf{p}_i = (\mathbf{n}-1) \times \mathbf{M} + \mathbf{m}}$ is satisfied.

With our example (**4**), three *a priori* valid combinations satisfy the equation. They are: (101, 101, 101), (73, 115, 115) and (87, 101, 115), which correspond to the two genera *EA–SP–(CH_3, CH_3, C_4H_9), *EA–SP–(CH_3, C_2H_5, C_3H_7) and to the subgenus *EA–SP–(C_2H_5, C_2H_5, C_2H_5). It should be noted that for all polyvalent ether superatoms the genera are built *without reference to the mass spectrum*. This is not the case for the two polyvalent alcohol superatoms *EA–S* and *EA–T*. Since α-cleavage plays a more important role for alcohols than for ethers, the program performs a preselection by constructing only the subgenera for which α-cleavage leads to a set of ions exhibiting a sum of intensities larger than 20%. With our example, from the three possible subgenera *EA–T(CH_3, CH_3, C_4H_9), *EA–T(CH_3, C_2H_5, C_3H_7) and *EA–T–(C_2H_5, C_2H_5, C_2H_5), only the first one is generated (see mass spectrum, Diagram 4).

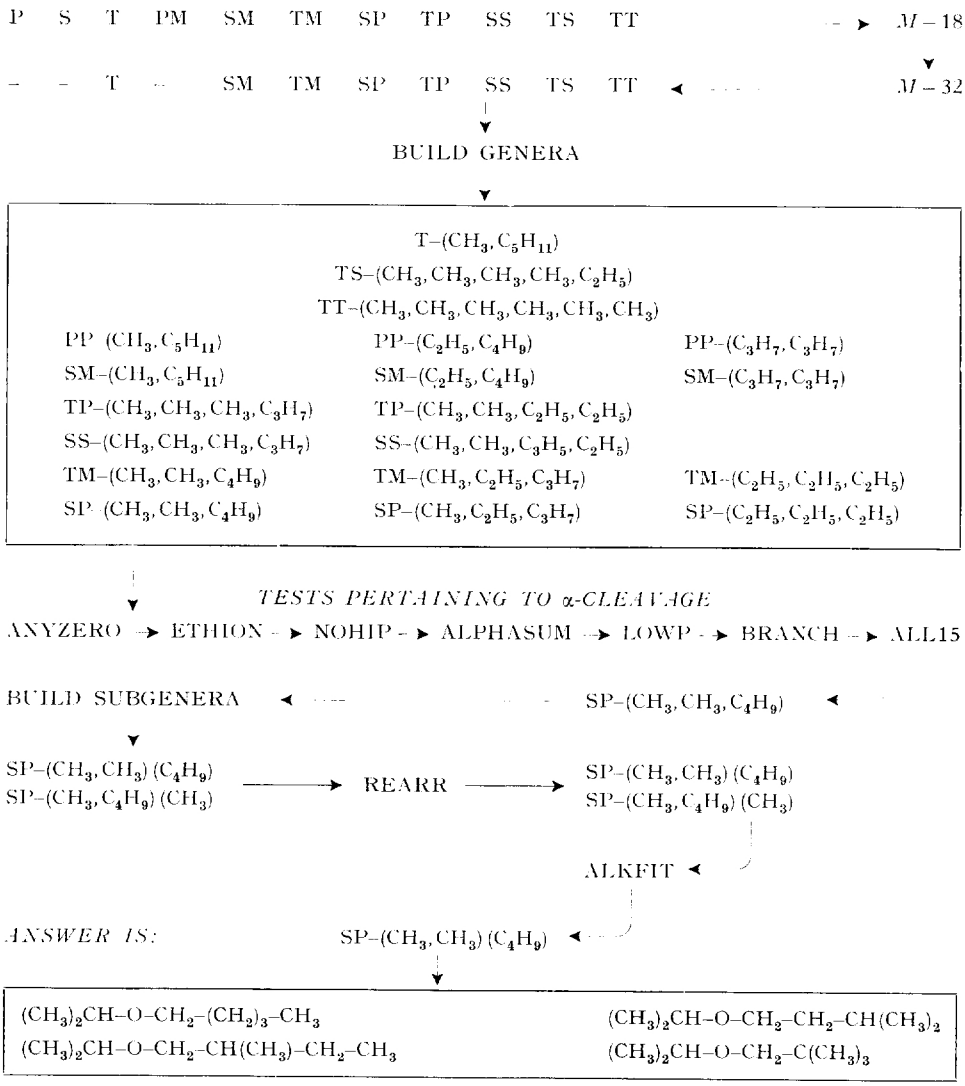
The validity of each genus is then tested for consistency with the mass spectrum. All the conditions about α-cleavage are included in the multistep test 'ALPHA' reported in Table 1. Diagram 5 illustrates how the program arrived at the correct solution for the mass spectrum of **4**. It shows which superatoms were discarded even before genera were constructed, which genera were built and how they were eliminated. All the subtests included in the general test 'ALPHA' are also recorded in Diagram 5. First the program requires that no potential α-fission peak except the $M-15$ peak be absent from the spectrum (subtest 'ANYZERO', Diagram 5). As there are no peaks corresponding to the loss of either C_2H_5 or C_3H_7 from the molecular weight in the mass spectrum of **4**, all genera with an ethyl or a propyl group attached to an α-carbon are eliminated. Out of the 19 genera and subgenera reported in Diagram 5, 13 were eliminated by that test. The next test is only performed for ethyl ethers having superatom *EA–PP* as a central subunit. For such subgenera the program requires that the ion CH_3CH_2OH (*m/e* 46) give a signal with an intensity greater than 2%. The subgenus *EA–PP–(CH_3, C_5H_{11}) is eliminated from further consideration by that test (subtest 'ETHION', Diagram 5).

Important α-series peaks (CH_2–XH_{v-1} + i × 14), having masses smaller than the mass of the ion arising from α-cleavage expulsion of the largest alkyl fragment, cannot be accounted for if the molecular ion is one not susceptible to undergo a favored rearrangement process according to the mechanism depicted under **2** (see test 'REARR' off in Table 1). Since *m/e* 45 is one of the major peaks in the mass spectrum of **4**, the two subgenera *EA–SM–(CH_3, C_5H_{11}) and *EA–TM–(CH_3, CH_3, C_4H_9) are rejected (subtest 'LOWP', Diagram 5). By the same reasoning the program will eliminate any molecule if the mass spectrum under study exhibits a strong signal (> 10%) at a mass value above that of the ion formed by α-cleavage expulsion of the smallest alkyl fragment (subtest 'NOHIP', Diagram 5). With our example all the remaining candidates contain at least one α-carbon bound methyl radical; since *m/e* 115 is the last peak in the mass spectrum, none of them is eliminated by that test.

Oxygen containing fragments formed by α-cleavage, even if they do not stabilize the positive charge as well as nitrogen containing ones do, still can compete with alkyl radicals for charge retention; this affords diagnostically useful ions, especially when the size of the alkyl group is not large enough to allow them to be highly branched.

Before performing the next test, the program checks the size of the biggest α-carbon bound alkyl group. If it is larger than C_3H_7 it could contain a quaternary carbon atom and would then favorably compete with the heteroatom for charge retention. In such a case, no minimum value is assigned to the sum of the intensities of the α-cleavage ions. Yet, if the α-carbon atoms of ether and thioether molecules

Diagram 5. *Description of the inference phase with isopropyl* n-*amyl ether* (**4**)



bear only small radicals, the total ion current carried by the α-cleavage ions should amount to at least a value representing 10% of the current carried by the ion giving the strongest signal. None of the remaining molecules were eliminated by that test (subtest 'ALPHASUM', Diagram 5). At 70 eV the larger alkyl fragment is preferentially expelled in an α-cleavage. For molecules which can generate more than one ion by α-cleavage, the program requires that each ion produced in such a way gives a stronger signal than the immediate next heavier ion formed by the same process, provided the alkyl radical expelled to give the heavier ion is smaller than C_3H_7. If it is C_3H_7 or larger, it could be a secondary or even a tertiary radical, and the program weakens its requirement; in such a case the intensity of the low mass ion has to be greater than $(0.5 + (0.1 \times \Delta\mathbf{C}) \times \mathbf{1})$ where **1** stands for the intensity of the higher mass ion and $\Delta\mathbf{C}$ for the difference in size between the two alkyl radicals lost to give the two α-cleavage peaks which the program compares. This test takes into account the possibility of branching as well as the respective sizes of the C_nH_{2n+1} radicals expelled (subtest 'BRANCH', Diagram 5). Candidate *EA–T–(CH_3,C_5H_{11}) is expected to give a stronger signal for ion $M-C_5H_{11}$ than for ion $M-CH_3$; since this is not the case in the mass spectrum tabulated in Diagram 4 (see *m/e* 115 and *m/e* 59), that molecule is rejected.

When a molecule has a methyl radical attached to one of its α-carbon atoms, the $M-15$ ion is often missing from the mass spectrum, especially when larger radicals can be expelled by α-cleavage from other sites. But, if all α-cleavages lead to the $M-15$ ion, *i.e.* if the molecule bears only methyl groups on its α-carbons, the program will keep such a molecule for further test only if the $M-15$ ion appears in the spectrum with a relative abundance exceeding the value of $20 \times (1-1/\mathbf{m})$, where **m** represents the number of methyl radicals attached to α-carbon atoms. The subgenus *EA–TT–(CH_3,CH_3,CH_3,CH_3,CH_3,CH_3) would have passed that test (subtest 'ALL15', Diagram 5) if *m/e* 115 had shown up with an intensity greater than $20 \times 5/6$, *i.e.* greater than 16%.

For all the remaining candidates for which there exists more than one way to distribute the alkyl radicals among the free valences of the superatom, the program then builds subgenera out of the genera. From *EA–SP–(CH_3,CH_3,C_4H_9), the only genus not rejected at that stage of the validation process, the program builds the two subgenera *EA–SP–(CH_3,CH_3)(C_4H_9) (**5**) and *EA–SP–(CH_3,C_4H_9)(CH_3) (**6**).

5

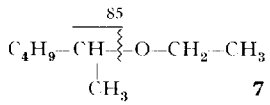
6

The program then simulates for structures **5** and **6** the rearrangement process depicted under **2**; it calculates the mass for every potential ion arising from such a mechanism. If at least one signal corresponding to such an ion is present in the mass spectrum with a relative abundance above 25% (15% for thioethers and 30% for amines), the molecule passes the test successfully (test 'REARR', Table 1). From structure **5** the two ions $CH_3-CH=\overset{+}{O}H$ (*m/e* 45) and $CH_2=\overset{+}{O}H$ (*m/e* 31) can originate from α-cleavage followed by simultaneous hydrogen transfer to the oxygen atom and C–O bond scission. Structure **6** can also lead to these two ions and, in addition, to ion

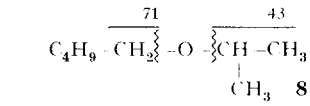
C_4H_9–CH=$\overset{+}{O}$H (*m/e* 87). Since *m/e* 45 has an intensity of 30% in the mass spectrum under study, both structures **5** and **6** are accepted by that test (see Diagram 5).

The last test which each remaining candidate undergoes is depicted in Table 1 as 'ALKFIT'. The final decision about keeping or rejecting a molecule depends on the relative abundances of the C_nH_{2n+1} ions formed by rupture of the C–O bond. The minimum intensity each alkyl ion should exihibit is related to its size and to the degree of substitution of the carbon atom which was originally an α-carbon of the molecular ion. The higher the degree of substitution of this carbon atom, the more likely is C–O bond rupture with charge retention on the alkyl moiety. Moreover, as large alkyl ions tend to further decompose, the bigger the alkyl ion the less important its diagnostic value as a potential ion.

The program requires that all C_nH_{2n+1} ions (with $\mathbf{n} > 2$) formed by cleavage of the C–O bond give signals with intensities exceeding the integer value of $(500 + (150 \times s))/\mathbf{n}^3$, where s represents the degree of substitution of the α-carbon atom (0 for $-\overset{|}{C}H_2$, 1 for $-\overset{|}{C}H-$ and 2 for $-\overset{|}{\underset{|}{C}}-$) and **n** the number of carbon atoms in the alkyl ion. Alkyl radicals which are branched at the α-carbon atom are thus required to yield stronger signals than the corresponding unbranched ones; the minimum required intensity decreases also as the size of the alkyl radical increases. For example, the relative abundance of a peak corresponding to a C_5H_{11} ion must exceed 4% if the α-carbon atom is not branched, 5% if it is mono-substituted, and 6% if it is di-substituted[18]; the above mentioned formula allows unbranched C_nH_{2n+1} ions to be missing from the mass spectrum if they are larger than C_7H_{15}. With our example, candidate **6** would have passed that test if at *m/e* 85, which corresponds to a C_6H_{13} ion formed according **7**,



7



8

a peak had been present with a relative abundance greater than $(500 + (1 \times 150))/216$, *i.e.* 3%. The correct molecule (**5**) is accepted by that final test. Peaks at *m/e* 43 (C_3H_7) and *m/e* 71 (C_5H_{11}) originating from the following cleavages (**8**) are bigger than respectively $(500+150)/27$, *i.e.* 24%, and 500/125, *i.e.* 4% (see spectrum tabulated in Diagram 4). Finally, a subroutine program calculates the number of isomers which are compatible with the structure of each subgenus inferred.

Diagram 4 shows that the program first selected from the mass spectrum of **4** a molecular weight of 116 amu., and henceforth attempted to validate a $C_7H_{16}O$ structure. Since no such molecule could fully explain the mass spectrum, the program repeated the process with $C_8H_{18}O$ and found the correct answer. The fact that the program did not get misled by the absence of the molecular ion at *m/e* 130 brings up the following question: *Would an experienced mass spectrometrist have rejected all $C_7H_{16}O$ isomers?*

The results we have obtained with 210 mass spectra are reported in Tables 2, 3, and 4. Results for 31 amine mass spectra other than the ones listed in Table 4

[18] These values are calculated from the formula $(500+(150+s))/n^3$, with $n = 5$ and $s = 0$, 1 and 2 respectively.

are already reported in one of our publications [3]. The correct structure is always included in the answer. In all cases the initial search space[19]) is already curtailed tremendously by using only mass spectral data. The results we have obtained

Table 2. *Results for ether and alcohol mass spectra*

Alcohol	Number of $C_nH_{2n+2}O$ isomers	Number of inferred isomers A	B	*Ether*	Number of $C_nH_{2n+2}O$ isomers	Number of inferred isomers A	B
n-butyl	7	2	1	Methyl *n*-propyl	7	2	1
isobutyl	7	2	1	Methyl isopropyl	7	3	1
sec-Butyl	7	3	2	Methyl *n*-butyl	14	2	1
2-methyl-2-butyl	14	1	1	Methyl isobutyl	14	2	1
1-pentyl	14	4	1	Ethyl isopropyl	14	1	1
3-pentyl	14	1	1	Ethyl *n*-butyl	32	4	1
2-methyl-1-butyl	14	4	2	Ethyl isobutyl	32	4	2
2-pentyl	14	2	1	Ethyl *sec*-butyl	32	2	2
3-hexyl	32	2	1	Ethyl *t*-butyl	32	1	1
3-methyl-1-pentyl	32	8	4	Di-*n*-propyl	32	1	1
4-methyl-2-pentyl	32	4	1	Di-isopropyl	32	1	1
1-hexyl	32	8	1	*n*-Propyl *n*-butyl	72	2	1
3-heptyl	72	4	1	Ethyl *n*-pentyl	72	4	1
2-heptyl	72	8	1	Methyl *n*-hexyl	72	8	1
3-ethyl-3-pentyl	72	1	1	Isopropyl *sec*-butyl	72	3	2
2,4-dimethyl-3-pentyl	72	3	1	Isopropyl *n*-pentyl	171	4	1
1-heptyl	72	17	1	*n*-Propyl *n*-pentyl	171	4	1
3-methyl-1-hexyl	72	17	6	Di-*n*-butyl	171	3	1
1-octyl	171	39	1	Isobutyl *t*-butyl	171	2	1
3-octyl	171	8	1	Ethyl *n*-heptyl	405	34	1
2,3,4-trimethyl-3-pentyl	171	3	1	*n*-Butyl *n*-pentyl	405	8	1
				Di-*n*-pentyl	989	10	1
1-nonyl	405	89	1	Di-isopentyl	989	18	7
2-nonyl	405	39	1	Di-*n*-hexyl	6 045	125	2
1-decyl	989	211	1	Di-*n*-octyl	151 375	780	1
6-ethyl-3-octyl	989	39	9	Bis-2-ethylhexyl	151 375	780	21
3,7-dimethyl-1-octyl	989	211	41	Di-*n*-decyl	11 428 365	22 366	1
1-dodecyl	6 045	1 238	1				
2-butyl-1-octyl	6 045	1 238	25				
1-tetradecyl	38 322	7 639	1				
3-tetradecyl	38 322	1 238	1				
1-hexadecyl	151 375	48 865	1				

A = Inferred isomers when only mass spectrometry is used.
B = Inferred isomers when the number of methyl radicals is known from NMR. data.

[19]) Since the program starts without knowing the elemental composition, it is not possible to assign a definite value to the size of the search space. Once the program has inferred an empirical formula $C_nH_{2n+v}X$ (v = valence of X), the search space includes all the isomers of empirical formulae $C_nH_{2n+v}X$, $C_{n-1}H_{2n+1+v}X$ and $C_{n+2}H_{2n+2+v}X$. The number of *a priori* possible isomers reported in tables 2, 3 and 4 for each compound, has been limited to all the isomers corresponding to the correct empirical formula. These numbers are calculated by a subroutine of the INFERENCE MAKER program. In one of our previous publications [8] the number of isomers with empirical formulae $C_{11}H_{24}O$ and $C_{12}H_{26}O$ have been wrongly reported to be 2460 and 6123 respectively; they should be corrected to 2426 and 6045.

also show that if NMR. spectra were used (only as methyl counters) the structure determination would be completely solved for many of the examples reported in Tables 2, 3, and 4.

It can be concluded, that even without the aid of NMR. spectrometry, the efficiency of the INFERENCE MAKER program is such that the PREDICTOR program of Heuristic DENDRAL cannot further differentiate between the inferred structures. If desired, the STRUCTURE GENERATOR program can be used to draw the structures. Although we agree that 'saturated acyclic monofunctional' molecules

Table 3. *Results for thioether and thiol mass spectra*

Thioether	Number of $C_nH_{2n+2}S$ isomers	Number of inferred isomers A	B	*Thiol*	Number of $C_nH_{2n+2}S$ isomers	Number of inferred isomers A	B
Methyl ethyl	3	1	1	*n*-Propyl	3	2	1
Methyl *n*-propyl	7	1	1	Isopropyl	3	1	1
Methyl isopropyl	7	2	1	*n*-Butyl	7	3	1
Di-ethyl	7	1	1	Isobutyl	7	3	1
Methyl *n*-butyl	14	3	1	*t*-Butyl	7	1	1
Methyl isobutyl	14	5	2	2-methyl-2-butyl	14	1	1
Methyl *t*-butyl	14	1	1	3-methyl-2-butyl	14	2	1
Ethyl isopropyl	14	1	1	3-methyl-1-butyl	14	6	3
Ethyl *n*-propyl	14	2	1	1-pentyl	14	4	1
Ethyl *n*-butyl	32	3	1	3-pentyl	14	5	3
Ethyl *t*-butyl	32	1	1	2-pentyl	14	6	3
Ethyl isobutyl	32	3	2	1-hexyl	32	8	1
Di-*n*-propyl	32	2	1	2-hexyl	32	12	5
Methyl *n*-pentyl	32	10	1	2-methyl-1-pentyl	32	8	4
Di-isopropyl	32	1	1	4-methyl-2-pentyl	32	4	2
Ethyl *n*-pentyl	72	4	1	3-methyl-3-pentyl	32	1	1
n-Propyl *n*-butyl	72	5	1	2-methyl-2-hexyl	72	8	3
Isopropyl *n*-butyl	72	5	2	1-heptyl	72	17	1
Isopropyl *t*-butyl	72	1	1	2-ethyl-1-hexyl	171	39	9
n-Propyl isobutyl	72	3	2	1-octyl	171	39	1
Isopropyl *sec*-butyl	72	4	3	1-nonyl	405	89	1
n-Propyl *n*-pentyl	171	4	1	1-decyl	989	211	1
Ethyl *n*-hexyl	171	8	1	1-dodecyl	6 045	1 238	1
Di-*n*-butyl	171	5	1				
Di-*sec*-butyl	171	3	1				
Di-isobutyl	171	3	1				
Methyl *n*-heptyl	171	21	1				
Di-*n*-pentyl	989	12	1				
Di-*n*-hexyl	6 045	36	1				
Di-*n*-heptyl	38 322	153	1				

A = Inferred isomers when only mass spectrometry is used.
B = Inferred isomers when the number of methyl radicals is known from NMR. data.

represent only a small fraction of all known organic compounds, it is interesting to realize that with those compounds, *the program in general performs better than an experienced mass spectrometrist.* More important perhaps is the fact that this kind of

research requires a formalization of mass spectrometry rules; such a formalization did not exist before. In view of the success with which the mass spectra of SAM compounds were interpreted, especially those of ethers and alcohols which are known to be difficult to interpret without taking advantage of low voltage data [9], we believe that no major obstacle exists which would prevent such a program from working with more complicated molecules.

Table 4. *Results for amine mass spectra*

Amine	Number of $C_nH_{2n+3}N$ isomers	Number of inferred isomers A	B	*Amine*	Number of $C_nH_{2n+3}N$ isomers	Number of inferred isomers A	B
1-propyl	4	1	1	N-methyl-di-isopropyl	89	15	3
Isopropyl	4	2	1	1-octyl	211	39	1
1-butyl	8	2	1	Ethyl-1-hexyl	211	24	1
Isobutyl	8	2	1	1-methylheptyl	211	34	1
sec-Butyl	8	4	2	2-ethylhexyl	211	39	9
t-Butyl	8	3	1	1,1-dimethylhexyl	211	32	4
Di-ethyl	8	3	1	Di-1-butyl	211	24	1
N-methyl-*n*-propyl	8	4	1	Di-*sec*-butyl	211	33	8
Ethyl-*n*-propyl	17	5	1	Di-isobutyl	211	17	5
N-methyl-di-ethyl	17	4	1	Di-ethyl-*n*-butyl	211	17	3
1-pentyl	17	4	1	3-octyl	211	26	2
Isopentyl	17	4	2	1-nonyl	507	89	1
2-pentyl	17	2	1	N-methyl-di-*n*-butyl	507	13	1
3-pentyl	17	5	1	Tri-1-propyl	507	2	1
3-methyl-2-butyl	17	4	1	Di-1-pentyl	1 238	83	1
N-methyl-1-butyl	17	4	1	Di-isopentyl	1 238	109	16
N-methyl-*sec*-butyl	17	3	1	N,N-dimethyl-2-ethylhexyl	1 238	156	9
N-methyl-isobutyl	17	4	1	1-undecyl	3 057	507	1
1-hexyl	39	8	1	1-dodecyl	7 639	1 238	1
Tri-ethyl	39	2	1	1-tetradecyl	48 865	10 115	1
2-hexyl	39	8	1	Di-1-heptyl	48 865	646	1
Di-1-propyl	39	8	1	N,N-dimethyl-1-dodecyl	48 865	4 952	1
Di-isopropyl	39	8	1	Tri-1-pentyl	124 906	40	1
N-methyl-1-pentyl	39	8	1	Bis-2-ethylhexyl	321 988	2 340	24
N-methyl-isopentyl	39	8	2	N,N-dimethyl-1-tetradecyl	321 988	3 895	1
Ethyl-*n*-butyl	39	6	1	(Di-ethyl)-1-dodecyl	321 988	2 476	1
N,N-dimethyl-1-butyl	39	10	1	1-heptadecyl	830 219	124 906	1
1-heptyl	89	17	1	N-methyl-bis-2-ethylhexyl	830 219	2 340	24
Ethyl-1-pentyl	89	16	1	1-octadecyl	2 156 010	48 865	1
1-Butyl-isopropyl	89	11	1	N-methyl-1-octyl-1-nonyl	2 156 010	15 978	1
4-methyl-2-hexyl	89	16	4	N,N-dimethyl-1-octadecyl	14 715 813	1 284 792	1

A = Inferred isomers when only mass spectrometry is used.
B = Inferred isomers when the number of methyl radicals is known from NMR. data.

Financial assistance from the *Advanced Research Projects Agency* (contract SD-183), the *National Aeronautics and Space Administration* (grant NGR-05-020-004) and the *National Institutes of Health* (grants AM-12758 and AM 04527) is gratefully acknowledged.

Experimental. – The computer program described here runs on the IBM 360/67 computer at the Stanford Computation Center. It is written in the LISP programming language. The computer can interpret low resolution mass spectra at a rate of 20 spectra per minute. Mass spectra which had not been reported in the literature were recorded in our laboratory, some with a *Varian* MAT CH-4 mass spectrometer, others with an AEI MS-9 instrument.

BIBLIOGRAPHY

[1] *Y.M.Sheikh, A.Buchs, A.B.Delfino, G.Schroll, A.M.Duffield, C.Djerassi, B.G.Buchanan, G.L.Sutherland, E.A.Feigenbaum* & *J.Lederberg*, Org. Mass Spectrom., submitted for publication.

[2] *G.Schroll, A.M.Duffield, C.Djerassi, B.G.Buchanan, G.L.Sutherland, E.A.Feigenbaum* & *J.Lederberg*, J. Amer. chem. Soc. *91*, 7740 (1969).

[3] *A.Buchs, A.M.Duffield, G.Schroll, C.Djerassi, A.B.Delfino, B.G.Buchanan, G.L.Sutherland, E.A.Feigenbaum* & *J.Lederberg*, J. Amer. chem. Soc., submitted for publication.

[4] *H.Budzikiewicz, C.Djerassi* & *D.H.Williams*, 'Mass Spectrometry of Organic Compounds', pp. 100–101, Holden-Day, San Francisco 1967.

[5] *Op. cit.* [4], p. 300.

[6] *Op. cit.* [4], pp. 9–14.

[7] *Op. cit.* [4], p. 297.

[8] *J.Lederberg, G.L.Sutherland, B.G.Buchanan, E.A.Feigenbaum, A.V.Robertson, A.M.Duffield* & *C.Djerassi*, J. Amer. chem. Soc. *91*, 2973 (1969).

[9] *Op. cit.* [4], p. 231.