$MEREG *external-regno* is the external registry number of the molecule (any uniquely identifying character string known to the database, for example, CAS number)

Square brackets ([]) enclose optional parameters.

An embedded MOLfile (see Section 3) follows immediately after the $MFMT line.

The forms of a *reaction* identifier closely parallel that of a molecule:

$RFMT [$RIREG *internal-regno* [$REREG *external-regno*]] *embedded RXNfile*

$PCRXN [$RIREG *internal-regno* [$REREG *external-regno*]] *embedded CPSS*

RXNfile [CP]

$RIREG *internal-regno*

$REREG *external-regno*

where

$RFMT defines a reaction by specifying its description as a RXNfile and $PCRXN [CP] defines a reaction by specifying its description as a CPSS-style RXNfile

$RIREG *internal-regno* is the internal registry number (sequence number in the database) of the reaction

$REREG *external-regno* is the external registry number of the reaction (any uniquely identifying character string known to the database)

Square brackets ([]) enclose optional parameters

An embedded RXNfile (see Section 6) follows immediately after the $RFMT line, and an embedded CPSS-style RXNfile follows immediately after the $PCRXN [CP] line

7.3. Data-Field Identifier. The [*Data-field Identifier*] specifies the name of a data field in the database. The format is

$DTYPE *field name*

7.4. Data. Data associated with a field follows the field name on the next line and has the form

$DATUM *datum*

The format of *datum* depends upon the data type of the field as defined in the database. For example: integer, real number, real range, text, molecule regno.

For fields whose data type is "molecule regno", the *datum* must specify a molecule and, with the exception noted below, use one of the formats defined above for a molecular identifier. For example

$DATUM $MFMT *embedded MOLfile*

$DATUM $MEREG *external-regno*

$DATUM $MIREG *internal-regno*

In addition, the following special format is accepted

$DATUM *molecule-identifier*

Here, *molecule-identifier* acts in the same way as *external-regno* in that it can be any text string known to the database that uniquely identifies a molecule. (It is usually associated with a data field different from the *external-regno*.)

8. CONCLUSION

A series of chemical structure file formats built up from one or more connection table blocks have been described. These formats allow for the storage and transfer of chemical structure information used typically for search queries, individual structures, or entire databases. It is hoped that these file formats will see even wider use.[6]

REFERENCES AND NOTES

(1) The various CTfile formats have been programmed, tested, and documented by a large number of people at MDL over the years. Besides the authors of this paper, these include S. Anderson, J. Barstow, R. Blackadar, T. A. Blackadar, R. Briggs, R. E. Carhart, B. D. Christie, J. D. Dill, G. Freitas, R. J. Greenberg, A. J. Gushurst, D. Henry, R. Hofmann, D. Horner, A. Hui, T. E. Moock, D. G. Raich, J. Steele, W. T. Wipke, and K. Wiseman-Sleeter.

(2) Wipke, W. T.; Nourse, J. G.; Moock, T. Generic Queries in the MACCS System. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Hampshire, 1984; pp 167–178.

(3) Gushurst, A. J.; Nourse, J. G.; Hounshell, W. D.; Leland, B. A.; Raich, D. G. The Substance Module: The Representation, Storage, and Searching of Complex Structures. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 447–454.

(4) Moock, T. E.; Christie, B.; Henry, D. MACCS-3D: A New Database System for Three-Dimensional Molecular Models in Chemical Information Systems. In *Beyond the Structure Diagram*; Bowden, D., Mitchell, E. M., Eds.; Ellis Howard: New York, 1990; pp 42–49.

(5) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of Atom–Atom Mapping and Related Features in the Reaction Access System (REACCS). In *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: New York, 1988; pp 303–313.

(6) For more details and information on future changes, contact Affinity, Molecular Design Limited, 2132 Farallon Drive, San Leandro, CA 94577.

Computer-Aided Molecular Formula Determination from Mass, ^{1}H, and ^{13}C NMR Spectra

B. G. DERENDJAEV,* S. A. NEKHOROSHEV, K. S. LEBEDEV, and S. P. KIRSHANSKY

Novosibirsk Institute of Organic Chemistry, USSR Academy of Sciences, Novosibirsk, USSR

Received March 5, 1991

A computer-aided technique for the determination of the molecular formula of a compound by its mass, ^{13}C, and ^{1}H NMR spectra is suggested. Efficiency of the method has been verified on 81 "unknowns". It has been shown that in 89% of instances the requested formula is found among the top three candidates of a computer answer, and in 45% of instances the computer suggests a single formula.

The use of computer systems for structure elucidation of organic compounds from a spectral data set is generally based on a known or assumed molecular formula.[1-3] This information was obtained by additional experiments (high-resolution mass spectrometry, CHN analysis, chemical analysis, etc.) or is postulated by the researcher from the background of the sample.

In the context of our work on a spectral data analysis system,[4-10] we have developed software to determine molecular formulas directly from analysis of the most simple and ac-

0095-2338/92/1632-0255$03.00/0

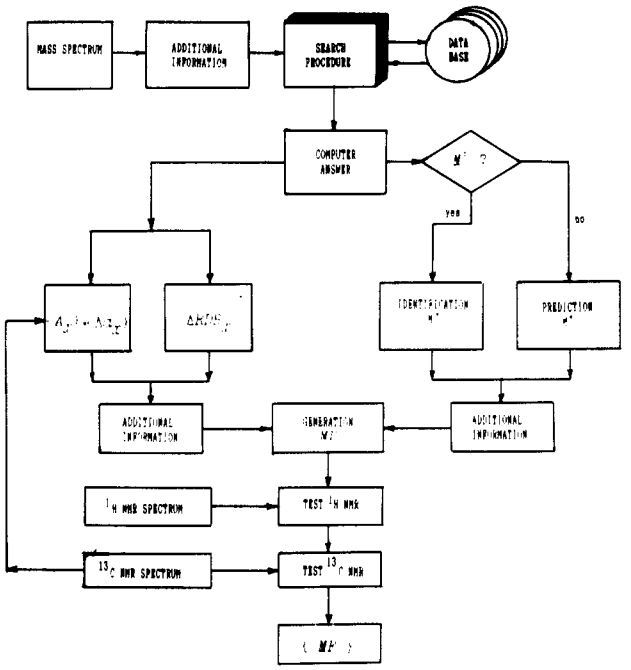

Figure 1. Functional scheme of determination of MF of unknown compound.

cessible spectral data. These are low-resolution electron impact mass spectra, the simplest ^{13}C NMR spectra with partial and complete spin–spin proton decoupling, and ^{1}H NMR spectra with position (or range) of chemical shifts and integrated intensities of signals. Such minimal spectral data requirements should make the software useful to more users. In this case it is unnecessary to provide detailed information about the molecular formula of a compound which is usually available from more sophisticated and expensive experiments such as multinuclear NMR spectroscopy or high-resolution mass spectrometry.

It seemed reasonable to develop an algorithm that would provide a ranked list of all molecular formulas consistent with the initial data and including the required formula at the top of a list. This would protect the researcher from making subjective and possibly erroneous decisions, and it is a reasonable goal of a computer-aided system.

The approach discussed below uses a mass spectral search procedure with a large database for determination of the molecular mass and formation of the initial list of possible molecular formulas of a compound from its mass spectrum, including those which have no molecular ion peaks. All components of the list of the possible molecular formulas are then checked for consistency with all available spectra or other data, if any, and are ranked. Figure 1 shows a general scheme of the approach.

Search Procedure. The procedure used is a part of the mass spectral interpretation system developed earlier.[9,10] The database of the system contains molecular masses, molecular formulas, chemical names, structures, and the spectra of ≈30000 organic compounds with molecular masses ranging from 26 to 702 amu.

The mass spectral search procedure selects from the database the compounds whose spectra are the best matches to the unknown spectrum according to the following criterion:

$$\mathrm{MF2} = 100WC/WX(\%)$$

where WC is a parameter specifying the significance of masses (m/z) and intensities (I) of matching peaks in the spectra; WX is a parameter characterizing the complete set of m/z and I values in the spectrum being analyzed. Details of this search algorithm are given in ref 10.

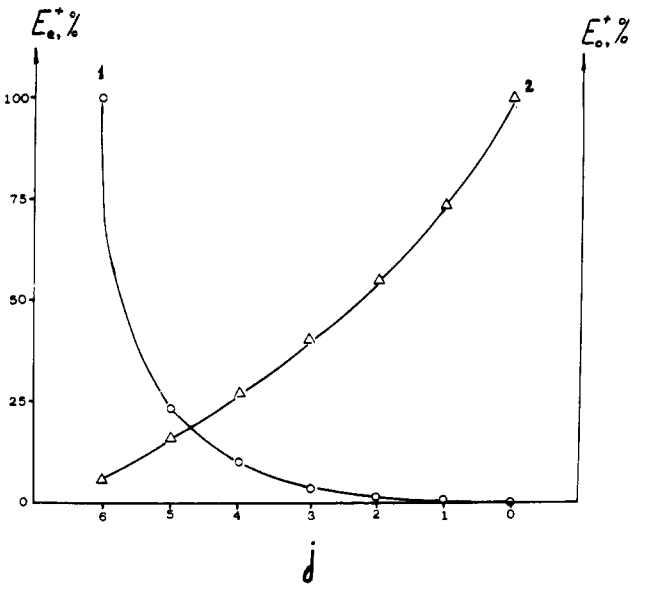

Figure 2. Plot of E_e^+ and E_0^+ parameters vs the j value.

Further analysis is performed on the information from the n first compounds of search results, using their spectra, molecular masses, and molecular formulas. The parameter n is assigned the value 3, 6, or 14, depending on the particular problem.

The establishment of the molecular formula (MF) of an unknown compound starts with determination of its molecular mass (MM) by its mass spectrum. There are several different ways of computer prediction of MM.[11-13] We divide the task into two parts:[4,6] molecular ion peak (M^+) identification and molecular mass prediction when an attempt to identify the molecular ion peak has failed.

M^+ Peak Identification. The software for M^+ peak identification uses traditional criteria [the maximal m/z value of ion peaks observed in the monoisotopic spectrum; masses of logical and impossible losses of neutral fragments from M^+] and two additional ones: the expected intensity (I_x^+) of the M^+ peak and the expected evenness (E_x^+) of the m/z value of the M^+ peak of the unknown compound (x). These parameters are determined from the results of the computer search. The I_x^+ value is the mean value of intensities (I_m^+) of M^+ peaks in the spectra of first three compounds of the computed answer.[4] The expected value of the evenness parameter (E_e^+ or E_0^+) is estimated from the ratio of occurrences among the first six compounds of the computed answer of j compounds with even or odd molecular mass values (Figure 2). For details, see refs 4 and 14.

All ions with m/z ranging from m_{max} to $m_{max} - 12$ amu are considered as possible M^+ ions where m_{max} is the mass number of ion peaks with the highest m/z value in a spectrum of the unknown, corrected for the natural isotopic abundance of ^{13}C. The M^+ candidates are ranked according to a parameter:

$$\mathrm{R1} = \alpha\beta\gamma(\mathrm{RI}_x^+ + E_x^+)/2$$

where α, β, and γ are coefficients equal to 0 or 1; $\alpha = 0$ if the M^+ candidate ion has primary neutral losses of 5–12 or 22–25 amu; $\beta = 0$ if $I_m^+ \leq 3\%$ of main peak intensity; $\gamma = 0$ if the candidate M^+ has an even m/z value and the first six compounds from the computer search have odd molecular mass values; and RI_x^+ is an empirical parameter between I_x^+ and the real intensity of M^+ found using the spectra of a teaching set.[4]

The larger the R1 value, the more reliable the given candidate. If none of candidate ions has R1 ≥ 10, the spectrum being analyzed is considered to have no molecular ion peaks.

Prediction of M^+. Instability of some molecules to electron

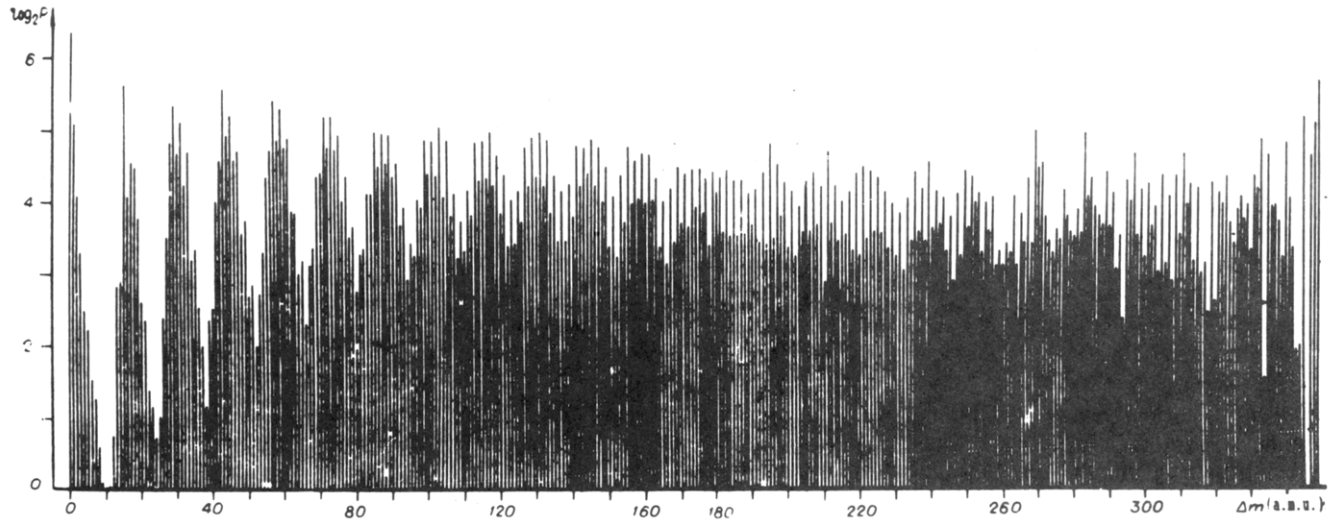

Figure 3. Plot of log $_2P$ vs primary loss values.

impact leads to the complete decomposition of M^+ in the ion source. As a result, such mass spectra show only fragmentation ion peaks, whose masses (m) differ from the mass of M^+ (M) by the Δm values characterizing the masses of neutral fragments eliminated from M^+ by the reaction:

$$M^+ \rightarrow m^+ + \Delta m$$

Hence, to determine the M^+ ion mass, one should know the Δm values that are added to fragmentation ion masses to give possible mass values of M^+ candidate ions.

For the generation of possible M^+ candidate ions, this approach uses one of the most intense ($I \geq 0.3\%$) fragmentation ion peaks (m_{max}) observed in the monoisotopic spectrum of the compound in the range of m/z from m_{max} to $m_{max} - 3$. The list of masses of neutral fragments is determined by analyzing monoisotopic spectra of the first 14 compounds from the search results:

$$\{\Delta m\},\ \Delta m_k = MM_n - m_k\ (0 \leq \Delta m_k \leq MM/2)$$

where MM_n is the molecular mass of the nth compound from the search (n = 1–14); $m_k = m/z$ of the kth peak in the spectrum of the nth compound from the search.

Simultaneously, for every Δm_k, the relative occurrence (p) of the given loss among the losses of n spectra of compounds of the search is calculated and stored. The list of M^+ candidate ions $\{M\} = \{M_1, ..., M_j, ..., M_s\}$ is obtained by adding the Δm_k values to m_{max}: $M_j = \Delta m_k + m_{max}$ ($0 \leq \Delta m_k \leq 100$). All the s elements of the $\{M\}$ set are molecular ion candidates.

For ranking we use the list of $\{\Delta m'\}_j$, $\Delta m'_l = (M - m_l)_j$, where m_l is the m/z value of the lth peak in the spectrum of the compound satisfying the relationship: $m_{max} \geq m_l \geq m_{max}/2$. The list elements ($\Delta m$) are then compared sequentially with the list elements $\{\Delta m'\}$. If the primary loss values are equal ($\Delta m_k = \Delta m'_k$), the total weight factor of the possible losses of molecular ion candidate is calculated:

$$WLY_j = \sum p^Y \log {}_2P^Y$$

where p^Y is the relative occurrence of a given primary loss in n spectra of the search; P^Y is the relative occurrence of the same loss among all database spectra. The histogram of the log $_2P^Y$ values is shown in Figure 3.

When $\Delta m_k \neq \{\Delta m'_l\}$, the total factor $WLN_j = \sum p^N \log {}_2P^N$ is calculated, where p^N and P^N are the same parameters as above, but characterizing the primary losses that are unlikely for the given M_j.

The final ranking of M_j mass value (j = 1, ..., s) is obtained according to the values of parameters

$$R2_j = (RL_j + E_j^+)/2(\%)$$

where $RL_j = WLY_j/(WLY_j + WLN_j)$. The larger the R2 value, the more reliable the given candidate.

Determination of Element Types, Number of Atoms, and Total Number of Double Bonds and Ring Equivalents. This problem is solved by analyzing the molecular formulas for six compounds of the results from the retrieval system. As a result of computer analysis, in the initial set of chemical elements $\{A^\circ\}$ = C, H, O, N, F, Cl, Br, S, P, I, and Si, such a subset $\{A_x\}$ may be identified, which describes the elemental composition of the unknown. A necessary condition for inclusion of an element from the $\{A^\circ\}$ list into the $\{A_x\}$ list is at least one occurrence of this element in the molecular formulas of the first six compounds of the computed answer. The use of such a simple approach, as shown earlier,[5] allows identification of $\{A_x\}$ subset elements with nearly 95% probability.

The probable number of atoms of the $\{A_x\}$ list elements is determined by the intervals $\Delta a_x = (a_{max}, a_{min})$, where a_{max} and a_{min} are the maximum and minimum numbers of atoms of this type in molecular formulas of the first six compounds of the search results.

The probable variations in the ring-plus-double-bond value of unknown are $\Delta RDB_x = (RDB_{max}, RDB_{min})$, where RDB_{max} and RDB_{min} are the maximum and minimum RDB values in the compounds (n = 6) of the computed answer, $RDB = a(4) + 0.5a(3) - 0.5a(1) + 1$, where $a(i)$ is the number of the atoms of i-valent element. The use of search system results for the estimation of the RDB values of unknowns has been described earlier.[15,16]

One can see from Figure 1 that at the stage of determination of element type and the number of element atoms, the software uses part of the ^{13}C NMR spectral data. Thus, the number of signals in the ^{13}C NMR spectrum may be used to correct the minimal number of carbon atoms (C_{min}) determined by the mass spectrum at the previous stage. The minimal number of hydrogen atoms (H_{min}) is calculated very simply from the multiplicity of the off-resonance spectrum. Simultaneously, the element composition of the compound is checked for the presence of oxygen atoms by the well-known ^{13}C NMR correlations. For example, the presence of a quartet in the range 48.5–60.0 ppm and a doublet or singlet in the range 167.0–220.0 ppm demands the presence of oxygen atoms in the $\{A_x\}$ list.

Generation, Control, and Ranking of Molecular Formulas. Molecular formulas are generated for all molecular masses {MM} of the list $\{M\}$, with the $\{A_x\}$, $\{\Delta a_x\}$, and ΔRDB_x data taken into account, to provide an exhaustive list of molecular formulas {MF} satisfying the conditions found. At the same stage, each of the molecular formulas generated is checked

for consistency with the ^1H NMR spectrum. If the number of signals in the ^1H NMR spectrum is more than one, the experimental value of relative integrals of the ^1H NMR spectrum must be matched with the number of hydrogen atoms of the given molecular formula on the assumption that the error of integration does not exceed 5%. Molecular formulas that do not match the ^1H NMR spectra are discarded from further consideration.

At the next stage of the algorithm (see Figure 1), the software again considers the ^{13}C NMR data. The system checks the possibility of complete decomposition of the hydrocarbon moiety of molecular formulas to CH_3, CH_2, and CH groups or their combinations identified from the sums of the same multiplicity signals in the ^{13}C NMR off-resonance spectrum. Solving the combinatorial problem is cut short as soon as at least one satisfactory decomposition is found. The corresponding molecular formula remains in the list of probable molecular formulas for further consideration. An additional filter checks the possibility of assigning the carbon and hydrogen atoms of each of the remaining molecular formulas to all signals of the ^{13}C NMR spectrum depending on the intensities of the latter.

The molecular formulas satisfying the above requirements are ranked using the RF1 or RF2 parameters, for the cases when the {*M*} list was obtained by identification of the M^+ or prediction of its mass number respectively. The RF1 and RF2 parameters are the products of R1 and R2 by the *Q* factor characterizing the statistical significance of separate elements of the molecular formulas among the first six compounds of the search result:

$$Q = v_{RDB} + \sum v_{ij}$$

where v_{RDB} is the occurrence of the given RDB value among these compounds; v_{ij} is the occurrence of *i*-type element with the number of atoms *j* in the same computer answer.

It can be seen that the software generates a ranked list of molecular formulas matching the spectra without using high-resolution mass spectrometry. This will also be formed when the M^+ are absent from the spectrum of an unknown. Certainly if the exact m/z value of M^+ is known, use of the well-known programs[17-19] will lead to a sharply reduced number of molecular formulas in the list. For this purpose other additional data about the compound, for example, CHN analysis data may be used.[1,20,21]

RESULTS AND DISCUSSION

We shall illustrate the work of this software with reference to computer analysis of the molecular formulas of two unknowns.

Example 1 (Problem 39.[22]) Figure 4 shows the mass spectrum and the ^{13}C and ^1H NMR spectra. The mass spectral search procedure selected compounds with the following molecular formulas: $C_7H_{14}O_2N_3$, $C_{10}H_{19}O_2N_1$, $C_9H_{17}O_2N_2$, $C_8H_{14}O_3N_2$, $C_{13}H_{22}O_3$, and $C_{14}H_{29}O_1N_1$. In accordance with the above algorithm, these formulas form a set $\{A_x\}$ containing element types C, H, O, and N. At a primary stage, the following range of C, O, N atoms numbers in the suggested molecular formula is established from their minimal and maximal numbers, with allowance for possible error of their determination[6] equal to ±2 units: C_{6-16}, O_{0-5}, and N_{0-5}. [The number of hydrogen atoms is calculated, but not established, in MF generation for the respective molecular mass values, considering the possible range of RDB. In the above case, RDB can vary from 0 to 4 unit.] It follows from the carbon magnetic resonance spectrum that $C_{min} = 6$, $O_{min} = 2$, and $H_{min} = 9$.

In accordance with the algorithm of molecular mass determination, the molecular ion peak was identified from the spectrum as MM = 171 (R1 = 70).

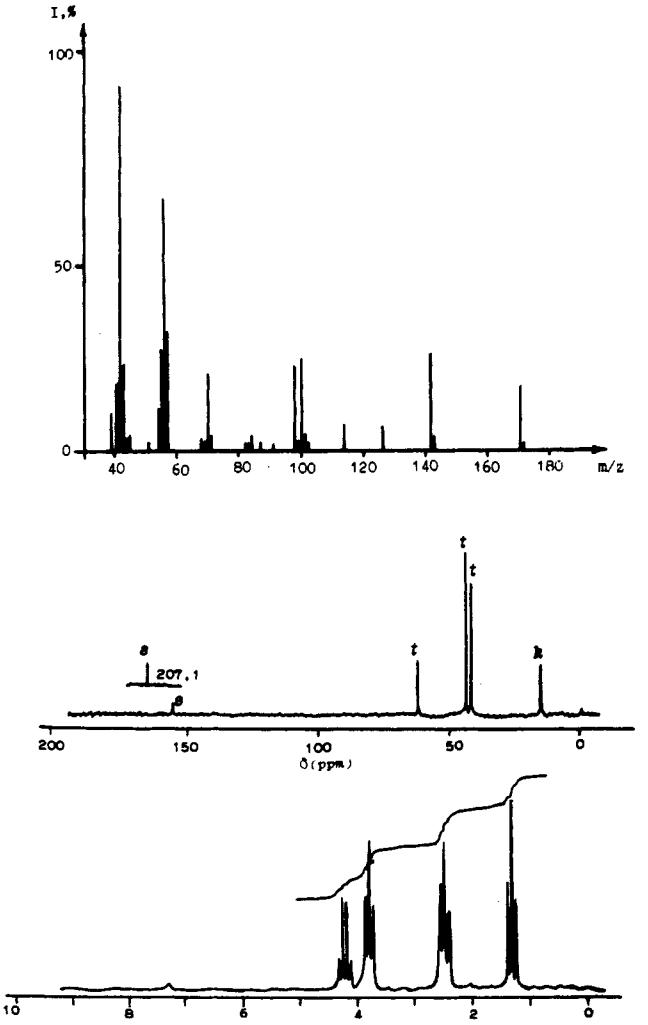

Figure 4. Mass, ^{13}C, and ^1H NMR spectra for problem 39.[22]

Chart I

Possible MM of Compound
MM-R1: /142-51/143-28/
Molecular Formulas of Selected Compounds

$C_9H_{21}N_1$ $C_9H_{21}N_1$ $C_9H_{21}N_1$ $C_8H_{16}O_2$ $C_{10}H_{20}O_2$ $C_6H_2O_6$
ΔA: C_{4-12} O_{0-8} N_{0-3}
ΔRDB: 0–7

Number of MF Generated: 16

MM	MF	RF	MM	MF	RF
142	$C_{10}H_{22}$	4590	142	$C_{11}H_{10}$	3060
142	$C_8H_{18}N_2$	2550	142	$C_8H_{14}O_2$	2550
142	$C_9H_{18}O_1$	2550	142	$C_7H_{10}O_3$	1530
143	$C_9H_{21}N_1$	1260	142	$C_6H_{10}O_2N_2$	1020
143	$C_{10}H_9N_1$	840	143	$C_8H_{17}O_1N_1$	700
143	$C_7H_{13}O_2N_1$	700	143	$C_7H_{17}N_3$	700
143	$C_6H_9O_3N_1$	420	143	$C_5H_9O_2N_3$	280
143	$C_6H_{13}O_1N_3$	0	142	$C_7H_{14}O_1N_2$	0

Number of MF after PMR Filter: 2
Number of MF After All Filters: 1

MM	MF	RF
142	$C_7H_{13}O_2N_1$	700

On the basis of these data, eight molecular formulas were generated. After the ^1H and ^{13}C NMR filters, however, only one molecular formula $C_8H_{13}O_3N_1$ remains. This formula is consistent with the data on the content of C, H, and N (56.2% C, 7.7% H, 8.2% N)[22] for the given example.

Example 2 (Problem 29.[22]) The spectra of another unknown are shown in Figure 5. Given in Chart I is a fragment of computer output from the MF determination using the spectra of an unknown.

As seen from the output, the molecular ion peak in this case has not been unambiguously determined. The computer has

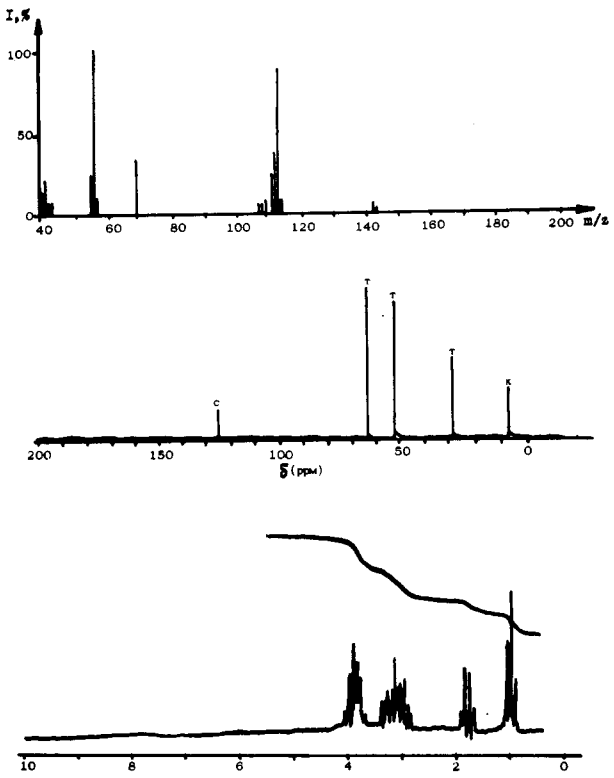

Figure 5. Mass, ^{13}C, and ^{1}H NMR spectra for problem 29.[22]

suggested two possible molecular masses of an unknown: 142 and 143 amu. The general list of MF generated for the given boundary conditions includes 16 items, but only two MF, $C_7H_{13}O_2N_1$ and $C_6H_{13}O_1N_3$, match the proton magnetic resonance spectrum. Reuse of the ^{13}C filter further reduced this list to a single molecular formula: $C_7H_{13}O_2N_1$. It is easy to see that it is consistent with the content values of C (58.6%), H (9.2%), and N (9.7%) given in ref 22 for this example, which were not used for MF determination.

The efficiency tests for this software have been carried out with mass spectra and ^{1}H and ^{13}C NMR spectra of individual compounds published in ref 22. The test list included the spectra of 81 compounds with molecular masses ranging from 72 to 346 amu with the general molecular formula: $C_{4-21}H_{5-45}O_{0-5}N_{0-4}S_{0-2}F_{0-1}$.

It was found that in 83% of instances, the true MF appears first in the lists. In 89% of cases, it was among the top three molecular formulas, and in 94%, among the top five molecular formulas. 45% of lists contained a single MF but about 3% of the lists did not include the true MF. On average, the lists of molecular formulas with the requested MF included 3.1 candidates. The number of candidates was generally larger when the molecular ion peak was not identified by the spectrum.

For illustration we shall give an example of analysis of the following spectral data (problem 57[22]):

mass spectrum: masses (intensities) 27 (110); 28 (120); 39 (60); 41 (170); 43 (650); 45 (50); 55 (55); 71 (180); 99 (40); 113 (30); 115 (1000); 116 (120)

^{1}H NMR spectrum: chemical shifts (intensities) 0.75–0.95 (30.0); 1.85–2.15 (5.0); 3.87–4.07 (10.0);

^{13}C NMR spectrum: chemical shifts (multiplicities) 17.5 (q); 35.0 (d); 67.8 (t); 117.0 (s). [q = quartet, t = triplet, d = doublet, s = singlet.]

In this example, the molecular ion peak was not identified, and probable molecular masses are represented by a list of 40 components.

The full initial list of molecular formulas generated contained 155 components, which were reduced by the filters to six molecular formulas. Molecular masses, molecular formulas, and their ranking parameters are given below:

MM	MF	RF	MM	MF	RF
130	$C_8H_{18}O_1$	805	174	$C_9H_{18}O_3$	81
142	$C_9H_{18}O_1$	400	174	$C_{13}H_{18}$	81
158	$C_9H_{18}O_2$	315	146	$C_8H_{18}O_2$	80

It is evident that only one molecular formula $C_9H_{18}O_2$ matches the content of C (68.4%) and H (11.5%) indicated for this compound.[22] Thus, in this and other cases which are difficult for analysis, the computer can also provide assistance for a researcher in molecular formula determination.

In Table I, one can follow the same general tendencies in the sequential reduction of the number of molecular formulas generated. The Table lists the molecular formula of an unknown, the number of molecular formulas generated using only mass spectral (MS) experimental data, the number of molecular formulas generated using the data on C_{min}, H_{min}, and O_{min} determined from the ^{13}C NMR spectrum (MS, ^{13}C NMR), the number of molecular formulas left in the list after the ^{1}H NMR of ^{13}C NMR filter and an additional ^{13}C NMR filter and the ranked list of molecular formulas in the respective computer answers. One can see that the total number of initially generated molecular formulas may be reduced by more than one order of magnitude. In every case, the efficiency of the list reduction depends on the type of spectral data.

Table I

MF of an unknown	no. of generated MF: MS	no. of generated MF: MS, ^{13}C NMR	no. of MF after filters: ^{1}H NMR	no. of MF after filters: ^{13}C NMR	MF candidates (range factor)
$C_4H_8O_1$	37	22	5	3	$C_4H_8O_1$ (5005); $C_4H_8O_2$ (372); $C_4H_8O_3$ (18)
$C_6H_{11}O_1N_1$	6	1	1	1	$C_6H_{11}O_1N_1$ (10440)
$C_{14}H_{18}O_5$	54	13	3	1	$C_{14}H_{18}O_5$ (3060)
$C_9H_{16}N_2$	7	4	3	1	$C_9H_{16}N_2$ (2000)
$C_{10}H_{10}O_2S_2$	48	43	10	5	$C_{10}H_{10}S_3$ (7932); $C_{10}H_{10}O_2S_2$ (4880); $C_9H_{10}O_3N_2S_1$ (2440); $C_9H_{10}O_1N_2S_2$ (2440); $C_{16}H_{20}O_1$ (1105)
$C_{18}H_{14}$	8	9[a]	5	3	$C_{18}H_{14}$ (4590); $C_{16}H_{10}N_2$ (3820); $C_{17}H_{10}O_1$ (2800)
$C_{19}H_{26}O_3$	11	3	2	1	$C_{19}H_{26}O_3$ (4000)
$C_8H_{13}N_1O_3$	12	8	3	1	$C_8H_{13}N_1O_3$ (4200)
$C_8H_{18}O_1$	46	25	9	4	$C_8H_{18}O_1$ (860); C_8H_{18} (800); $C_8H_{18}O_2$ (60): $C_8H_{18}S_1$ (40)

[a] Increased number of generated *MF* is explained by the reduced minimal number of carbon atoms given by the number of signals in the ^{13}C NMR spectrum.

REFERENCES AND NOTES

(1) Oshima, T.; Ishida, Y.; Saito, K.; Sasaki, S. *Anal. Chim. Acta* **1980**, *122*, 95.
(2) Gribov, L. A.; Elyashberg, M. E.; Koldashov, V. N.; Pletnjov, I. V. *Anal. Chim. Acta* **1983**, *148*, 159.
(3) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986.
(4) Nekhoroshev, S. A.; Derendjaev, B. G. *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk* **1985**, *1*, 101.
(5) Derendjaev, B. G.; Nekhoroshev, S. A. *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk* **1985**, *1*, 108.
(6) Nekhoroshev, S. A.; Kirshansky, S. P.; Derendjaev, B. G. *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk* **1985**, *6*, 113.
(7) Lebedev, K. S.; Tormyshev, V. M.; Derendjaev, B. G.; Koptyug, V. A. *Anal. Chim. Acta* **1981**, *133*, 517.
(8) Derendjaev, B. G.; Nekhoroshev, S. A.; Kirshansky, S. P.; Lebedev, K. S. *Zh. Anal. Khim.* **1987**, *42*, 1312.

(9) Kirshansky, S. P.; Lebedev, K. S.; Derendjaev, B. G. *Zh. Anal. Khim.* **1987**, *42*, 1092.
(10) Lebedev, K. S.; Kirshansky, S. P.; Nekhoroshev, S. A.; Derendjaev, B. G. *Zh. Anal. Khim.* **1987**, *42*, 1320.
(11) Jardine, A.; Reed, R. I.; Silva, M. E. S. F. *Org. Mass Spectrom* **1973**, *7*, 601.
(12) Dromey, R. G.; Buchanan, B. G.; Smith, D. H.; Lederberg, J.; Djerassi, C. *J. Org. Chem.* **1973**, *40*, 770.
(13) Mun, I. K.; Venkataraghavan, R.; McLafferty, F. W. *Anal. Chem.* **1981**, *53*, 179.
(14) Mun, I. K.; Venkataraghavan, R.; McLafferty, F. W. *Org. Mass Spectrom.* **1981**, *16*, 82.
(15) Dayringer, H. E.; McLafferty, F. W. *Org. Mass Spectrom.* **1977**, *12*, 53.
(16) Derendjaev, B. G.; Lebedev, K. S.; Nekhoroshev, S. A. *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk* **1981**, *2*, 91.
(17) Burlingame, A. L. *Adv. Mass Spectrom.* **1968**, *4*, 15.
(18) Gorfinkel, M. I.; Nekhoroshev, S. A. *Avtometrija* **1972**, *3*, 126.
(19) Dromey, R. G.; Foyster, C. T. *Anal. Chem.* **1980**, *52*, 394.
(20) Sukharev, Yu. N.; Nekrasov, Yu. S. *Zh. Anal. Khim.* **1981**, *36*, 2176.
(21) Fürst, A.; Clerc, J.-T.; Pretsch, E. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 329.
(22) Fuchs, Ph. L.; Bunnell, Ch. A. *Carbon-13 NMR based organic spectral problems*; John Wiley and Sons: New York, 1979.

COMPUTER SOFTWARE REVIEWS

MOBY. Version 1.41

NIKOLAI S. ZEFIROV* and IGOR I. BASKIN

Moscow State University, Russia

Received February 5, 1992

MOBY[1] is a program for molecular modeling. It can be run on IBM compatibles (with 640K of RAM, 80 × 87 arithmetic coprocessor, hard disk, VGA, EGA, or HERCULES graphics adapters). The use of a mouse is optional. Both English and German versions are available. The program is shipped on two 5.25-in. or 3.5-in. diskettes that are not copy-protected.

The program can be run in demo mode allowing the user to get acquainted with some of its capabilities. In protocol mode the whole session is captured to a disk file, which may be edited and shown in demo mode. A manual for the program contains a thorough description with an extensive tutorial.

MOBY allows the user to draw structures (up to 2000 centers and bonds) with a 3D editor, read it from disk (a number of formats including MNDO, Brookhaven Protein Database, MACCS II, SHAKAL, and freely definable format are supported), display the structure and its van der Waals or solvent surfaces, define a fragment (up to 150 centers), carry out force-field calculations (extended AMBER force field) with exact calculation of their interaction with all centers (molecular dynamics simulation, conformational analysis for up to six torsion angles, calculations under geometry constraints and with periodic boundaries are supported), and conduct semiempirical quantum chemical calculations with MNDO and AM1 methods for systems up to the size of glucose. Matching of structures and interactive docking are also provided.

It is worthwhile mentioning the flexibility of the force-field calculations. Hydrogen atoms bound directly to carbon atoms can be considered explicitly or implicitly ("united atom" approximation). Either all terms or only the valence terms of the force field can be used for the energy calculations. Geometry optimization can be run with constraints specified by the user for interatomic distances, bond lengths, valence, and torsion angles. For special cases periodic boundaries can also be specified. The conjugate gradient and the steepest descent optimizer can be selected, and various options and criteria for geometry optimization can be defined. Easy switching between force-field and quantum chemical calculations provides the ability to use charges calculated by the AM1 method for molecular mechanics calculations and to use the geometry optimized by the force-field method for quantum chemical calculations.

The main feature distinguishing MOBY from many other molecular modeling programs on PCs is its orientation toward biochemistry. The program uses the AMBER force field,[2] known to be well-suited for proteins and other biological molecules. Protein structures can be read directly from the Brookhaven Database. The program displays the protein and any of its parts in many different ways (all atoms and bonds, backbone, sequence of amino acid residues, etc.). The user can define a fragment in the protein and investigate its conformational behavior. Examples in the tutorial illustrate the modeled denaturation of a polyalanine helix in cytochrome C and substrate binding to the enzyme adenylate kinase. The coloring ability is a very interesting feature of the program. All centers can be colored in accordance with their properties (charge, energy, etc.). This makes the process of geometry optimization quite unusual in appearance.

MOBY can be used in various fields such as biochemistry, enzyme catalysis, and drug design, but it cannot be recommended for the use in conformational analysis in areas outside biochemistry because of the relatively "poor" parameterization of AMBER force field. In such cases programs like PCMODEL are preferable. We have tried to compare these two programs by calculating the conformational energy of a methyl group. PCMODEL (extended MM2 force field) gave 1.78 kcal/mol while MOBY gave 1.10 kcal/mol. The experimental value in aprotic solvent is 1.70 kcal/mol.

Also, it must be pointed out that it is not easy to use this program without having studied the manual or tutorial, since pleasant features of user-friendly programs such as pop-up helps or use of the ESC key for exiting are absent.

Overall, the program is recommended, especially for biochemistry and drug design.

REFERENCES AND NOTES

(1) MOBY is available from Springer-Verlag, New York, Inc., P.O. Box