each plant based on computerized research and ethnomedical information.

The end result was the identification of approximately 300 plants from a list of more 4500 that appeared to be the most promising for initial study. Fifty of these, chosen for their availability in the locality of a particular research center, have undergone preliminary investigation. Eight have provided confirmed desirable activity within two different laboratory animal assays. Although the true value of the NAPRALERT approach must await the successful development of clinical drugs from these plants, it would appear the utilization of such a computer-generated analysis can be an indispensible adjunct to natural products research. Certain aspects of this predictive program have been published elsewhere.[1,2]

Information contained in the NAPRALERT file has also been used by the National Cancer Institute, as well as by the herbal, pharmaceutical, and cosmetic industries in the development of new products. Future considerations for the use of the NAPRALERT-type database include the direct preparation of handbooks for the natural products researcher and the preparation of phylogenetic "density maps" for use in chemotaxonomic and biotaxonomic decisions, as well as the possible prediction of an impending endangered species through periodic examination of newly constructed maps.

These are but a few of the possibilities to which this comprehensive file of scientific data on natural products can be used. Individual needs and advanced computer technology will dictate future resource applications.

ACKNOWLEDGMENT

This work has been supported in part by the National Cancer Institute, Contract NCI-CM-97259, and the World Health Organization, Project 78135.

REFERENCES AND NOTES

(1) Soejarto, D. D.; Bingel, A. S.; Slaytor, M.; Farnsworth, N. R. "Fertility-Regulating Agents from Plants". *Bull. W. H. O.* **1978**, *56*, 343–352.

(2) Farnsworth, N. R.; Loub, W. D.; Soejarto, D. D.; Cordell, G. A.; Quinn, M. L.; Mulholland, K. "Computer Services for Research on Plants for Fertility Regulation". *Korean J. Pharmacogn.* **1981**, *12*, 98–110.

CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra[†]

HERMANN KALCHHAUSER and WOLFGANG ROBIEN*

Institut für Organische Chemie der Universität Wien, A-1090 Vienna, Austria

Received July 13, 1984

A computer program for the analysis of ^{13}C NMR spectra by various search strategies, including different methods for line search, molecular formula search, and structure-oriented search, is presented. The key algorithm of the program performs a fully automated assignment of ^{13}C NMR resonances to the respective carbons of a known structure. A database of 8000 ^{13}C NMR spectra taken from the literature and from our own measurements was created, containing carbon-centered substructural environments and their corresponding chemical shifts. The assignment algorithm is based on the prediction of chemical shift ranges from these data and permits a stepwise solution of the assignment problem with chemical shift arguments up to a five-bond radius.

INTRODUCTION

During the last decade NMR instrumentation became more sophisticated, and carbon-13 NMR data are now routinely reported in many papers dealing with natural product chemistry. The interpretation of ^{13}C NMR measurements is based on multiplicities, either from SFORD or *J*-modulated spectra, and on chemical shifts, which are mainly used to determine the number of sp^2 carbons and some functionalities with narrow shift ranges like methoxy groups. The chemical shift value of a certain carbon resonance depends strongly on the environment of the corresponding carbon. This sensitive probe cannot be fully utilized by manual interpretation of carbon-13 resonance data. The number of published reference data exceeds many thousand spectra per year; therefore, computerized databases have been built up.[1–16] In this paper we describe our program package, which includes spectrum estimation, many different file search strategies, and the complete automated assignment algorithm.

DATA STORAGE AND OVERALL DESIGN

Each reference data set contains the information given in Table I. From these data, several subfiles containing special information can be created by the computer allowing efficient execution of the different search strategies. The search methods available in the program are given in Table II. Each search function can be called by a three-letter abbreviation. A second program, named C13ADD, performs all other tasks concerning addition, modification, and control of existing records and also includes routines for generation of sorted lists by name, bibliography, or molecular formula. The CSEARCH program is designed for interactive use, but every search can be performed as batch job without user interaction.

Table I. Data Stored in Each Record

(1) entry number
(2) compound name, as given in the literature, up to 160 bytes
(3) comment and experimental conditions (temperature, reference, ...)
(4) structure: atom type, connectivity matrix, and bond type, up to $C_{40}H_{99}O_{63}$ and all other elements up to 15
(5) solvent
(6) molecular formula
(7) literature
(8) chemical shifts and multiplicities
(9) assignment of the resonance lines

SEARCH STRATEGIES

ISO. This option allows the user to find all compounds in the database having a specific molecular formula. During one

[†] Dedicated to Prof. Dr. K. Schlögl on the occasion of his 60th birthday.

Table II. Available Search Strategies

(a) searches using the molecular formula	
ISO:	search for isomers
MOF:	search by partial molecular formula
HOM:	search for a homologuous series
(b) searches using shift values	
LIN:	search for single lines
SUB:	search for groups of lines without multiplicity
GRO:	search for single lines or groups of lines using multiplicity (optical, combines LIN and SUB)
EQU:	SAHO search[17] for identical spectra
SIM:	SAHO search for similar spectra
SPH:	search for all possible functional groups[18] suitable to a certain line
(c) structure-oriented searches	
QUI:	spectrum estimation for a given structure
SPC:	same as QUI but, additionally, the identity and information on all reference spectra used are given
PAR:	substructure search
RAN:	substructure search combined with shift range calculation of defined carbons
RIN:	search for ring-size combinations
LIM:	shift range for a functional group
(d) assignment of ^{13}C NMR spectra	
ASS:	perform fully automated assignment procedure
(e) additional possibilities	
REF:	show a reference by its entry number
NAM:	search for compound names
LIT:	show the stored bibliography
PLO:	graphical output of structure and spectrum
NEW:	get information from the administrator
REM:	send a remark to the administrator
STA:	get statistical information about the database
HEL:	call help function
ADD:	add reference data; these data are only stored in the input format for later addition
ACC:	get accounting information
DEL:	rewind output file and rewrite header
COM:	insert header

search run up to 15 molecular formulas can be processed. The result is presented as a table containing the molecular formula and the number of hits found. The user can direct the output to either the terminal and/or the line printer.

MOF. The lower and upper limits for the molecular formula are defined, and all compounds fitting these limits are shown by reference number, compound name, and bibliography.

HOM. The molecular formulas for a homologuous series are calculated from a starting value and an increment. The series is constructed up to 40 carbons or 99 hydrogens or up to 15 atoms for all other elements. Afterward, a search by molecular formula, as described under ISO, is performed.

LIN. Up to 15 lines can be searched concurrently. Every reference that contains at least one of the defined lines fits the search condition. In order to speed up the search procedure, a bit pattern is created by dividing the shift range from 0 to 240 ppm into 240 divisions and setting the appropriate bit to 1, when a line occurs in the corresponding ppm range. Multiplicities from SFORD or *J*-modulated spectra can be used, or the multiplicity can be omitted. This search procedure can be only applied to very uncommon lines, since it produces too much output for common shift values.

SUB. From the given shift values the same bit string as described under LIN is generated and compared with the reference data. The number of common bits is counted and serves as a measure for spectral similiarity. Despite neglection of multiplicities, this option gave good results.

GRO. The GRO search combines LIN and SUB and is completely independent of the input sequence of the chemical shift values. This method is much more time consuming than a convergent search strategy but has the advantage that the result can be manipulated in many ways afterward by quick bit operations, which allow subtle feature selection. Up to 40 lines can be processed during one search. The result is stored in a two-dimensional bit array, the columns belonging to the defined lines and the bits within a row being numbered consecutively and set to 1 when the appropriate reference contains the specified line. The result consists of the number of references belonging to each line (=number of bits set to 1 within a row) and the number of references ordered according to the number of coincident lines (=number of bits set to 1 within a column). This matrix can be generated during a batch job and stored for interactive interpretation, because only matrix generation is a time-consuming task. Matrix manipulation can be performed easily, and certain constraints on the lines can be defined, including occurrence and absence of lines. After each input, the result is calculated and displayed. The user can specify further conditions or transfer this bit array to the output routine.

EQU and SIM. These options allow the user to perform a SAHO search as described by Bremser et al.[17] with slight modifications based on the word length of 60 bits of our computer equipment. In our case, the spectral region is divided into 20 12 ppm ranges, using 3 bits for each part. EQU performs a search for identical patterns and SIM for similar spectral patterns.

SPH. A very useful task for spectrum interpretation is to find all suitable functional groups belonging to one resonance line. The input data required are the shift value, the multiplicity, and the standard deviation. The output contains all functional groups fitting these limits ordered according to the number of their occurrence. For every functional group, the whole shift range, the mean shift value, and the total number of its occurrence are calculated (see LIM) to support the interpretation of the results.

QUI. This option performs a spectrum task for a given structure. The user can specify the number of shell levels (from 1 up to 5), which are included in the computation. Spectrum prediction is based on the HOSE code.[19] This carbon-centered substructural code defines the environment of each carbon and can be generated from the structural input via the connection table. A file containing these codes and their corresponding chemical shift values of all reference data is available and can be searched for the codes generated from the query. The stored HOSE codes of our database describe at least the complete five-bond environment of each carbon (hydrogen is incorporated into the atom symbol), which is extremely useful in spectrum estimation of aromatic compounds in order to include the influence of para substitution. Typical CPU time for spectrum calculation of a steroid with about 70 000 HOSE codes is in the range of 2–3 s.

SPC. This option performs the same task as QUI; additionally, the entry numbers of all used reference spectra are stored, and an identity search is done.

PAR. This task performs a conventional substructure search with much less screens (only 120) than other especially designed systems,[20–23] because this search strategy is not very often used and the database is small.

RAN. This search method performs also a substructure search as described under PAR; additionally, the user can specify carbons for which the shift range should be calculated.

RIN. The RIN search is used to retrieve all compounds with a user-specified combination of ring sizes. The references fitting the input conditions are shown by number, name, and bibliography.

LIM. The total shift range for a functional group, the mean shift value, and the number of hits found are displayed. This information is created during database update automatically and stored on a separate file.

NAM. The NAM command searches for user-defined compound names or name fragments. This is a string search, which is accelerated by a screening file. During database creation, a file is generated automatically that contains every chemical substance name encoded in a bit pattern, describing the occurrence of the letters within the name. This simple screen excludes in most cases about 80–90% of the reference data.

REF. This command shows a reference data set by its entry number.

PLO. This command calls a graphic package including spectrum display and the excellent structure display program of Shelley.[24]

LIT and ZIT. These commands call the stored bibliography and the corresponding references, respectively.

ACC. This command shows the users their accounting information including number of sessions, CPU time, and number of printed pages. The administrator gets this information for all users and, additionally, the number of applications and the average CPU time for each search method, a valuable tool for program optimization and further developments.

AUTOMATIC ASSIGNMENT OF CARBON-13 NMR SPECTRA

Assignment of carbon-13 NMR spectra is often a very tedious task and can be done in two different ways: (a) by comparison with literature data and/or (b) by independent experiments like selective decoupling or two-dimensional NMR spectroscopic techniques.[25] Usually, both methods are applied. The first one can be completely done by an appropriate computer program,[26,27] and the result of this computational procedure can help in planning further experiments in order to resolve remaining ambiguity. We have therefore developed a simple algorithm, which requires only the following input parameters: the known structure of the compound and the resonance positions of the carbons including multiplicities (optional) as given in the peak list from the NMR spectrometer.

In the first step of the computation, the connectivity table is generated from the input structure and converted to the HOSE codes by using a one- up to five-bond level. For these codes, the corresponding chemical shift values are looked up in the appropriate substructure file giving shift ranges to be expected for all carbons. These predicted shift ranges are extended depending on the standard deviation and the number of references found:

$$\mathrm{LL} = \mathrm{CS}_{\min} - X \qquad \mathrm{UL} = \mathrm{CS}_{\max} + X$$

If REF = 1, $X = 15/N$; if REF < 30–5N, X = 1.5SD; if REF ≥ 30–5N, X = SD. LL (UL) is the lower (upper) limit of the shift range; $\mathrm{CS}_{\min}$ ($\mathrm{CS}_{\max}$) is the lowest (highest) shift value predicted from the database; REF is the number of corresponding codes found; N is the bond level; SD is the standard deviation.

For a compound containing n carbons, a $n \times n$ matrix is created, and all elements are set to 1; values of 0 and 1 are allowed. If the element $M(m,n)$ is set to 1, line n is assigned to carbon m. The problem is now to reduce the number of non-zero elements within the matrix in such a way that every row and column contains exactly one non-zero element. First, the assignment procedure uses multiplicity information, if available. The next step includes application of the predicted shift ranges calculated at the one-bond level and consequent reduction of the non-zero matrix elements. This elimination proceeds in the same manner at the further bond levels. Between each assignment cycle, the matrix is controlled for physically meaningless combinations of non-zero elements. A resonance line cannot be definitely assigned to more than one carbon, and every line must be used. If such an error occurs, the algorithm tries to remove these inconsistencies, which leads to a further reduction of the matrix. If the method fails, which never occurred during 12000 calculations, the computation is stopped.

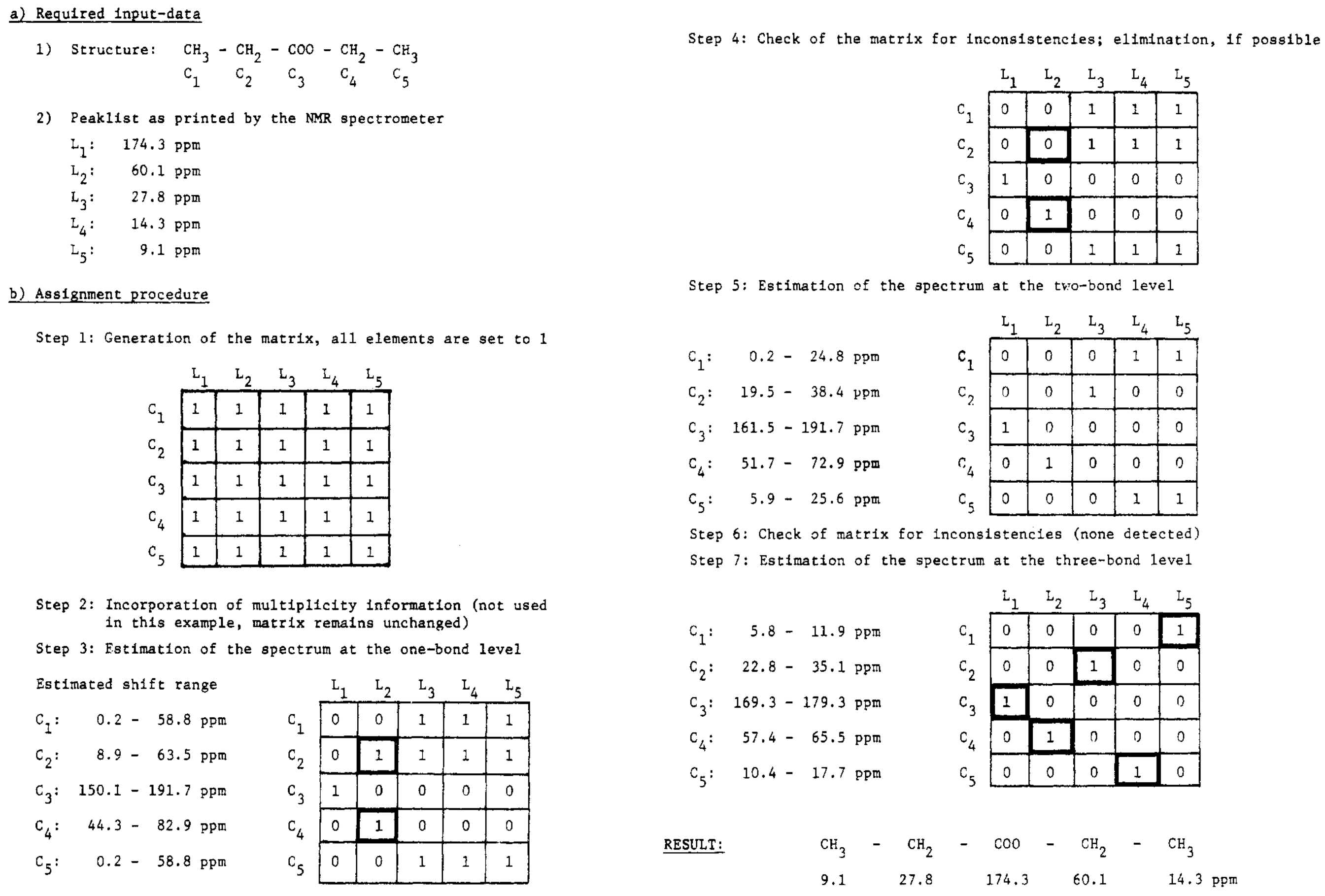
The detailed computational process is shown in Figure 1 for the ^{13}C NMR data of propionic acid ethyl ester. The assignment starts with the creation of a 5 × 5 matrix containing 25 non-zero elements. The assignment is complete when this number is reduced to 5 in such a way that every column and every row contains exactly one non-zero element. During the next step the shift ranges are calculated via the HOSE codes at the one-bond level. With these ranges, L_1 (174.3 ppm) is assigned to C_3, and L_2 (60.1 ppm) corresponds to C_2 or C_4. The three high-field lines cannot be distinguished at this step, and therefore, the problem is reduced from a 5 × 5 to a 3 × 3 matrix. Now the matrix is checked for impossible combinations of non-zero elements; such an inconsistency is detected. (see Figure 1, steps 3 and 4). In this case, line L_2 is definitely assigned to C_4 and therefore cannot be used within the row of C_2. Estimation of the spectrum at the two-bond level reduces the problem to a 2 × 2 submatrix; at the three-bond level both methyl groups can be assigned correctly. The procedure can also take in consideration the known assignment of some lines, which allows one to incorporate data from additionally performed NMR experiments in order to support the assignment process.

RESULTS

Example. The assignment of the ^{13}C NMR spectrum of 17β-hydroxy-1,4-androstadien-3-one with different multiplicity informations is shown in Table III. The computer-assisted assignment starts with the generation of a 19 × 19 matrix containing 361 non-zero elements. Using the full multiplicity information reduces this number to 105 (4 × 4 for the singlets, 7 × 7 for the doublets, 6 × 6 for the triplets, and 2 × 2 for the quartets). Application of the predicted shift ranges from the database at the one-bond level identifies the resonance lines of C-3, C-5, and C-17. At the three-bond level, a significant reduction of the non-zero elements to 33 occurs, and now C-1, C-8, C-12, C-18, and C-19 are correctly assigned. After application of the predicted shift ranges at the five-bond level, all lines within the sp^2 region are identified, which is especially remarkable for the pair C-2 and C-4. Within the sp^3 region, five lines are assigned to the corresponding carbons; the other ones can be divided into four groups of interchangeable lines. The shift differences within these groups are quite small; despite this fact, all lines can be further assigned from the estimated shift values from the calculation. In Table III the results are also given for J-modulated spectra; in this case the same result is obtained. Ommitting multiplicity information, an additional ambiguity occurs for the pair C-8 and C-12. These excellent results demonstrate the possibilities of our assignment algorithm.

A further example for the computer-assisted assignment is given in Table IV with the ^{13}C NMR data of roquefortine. The input structure is represented by the reference data only at the one- or two-bond level for most of the carbons. The estimated shift values for those carbons differ very much from the experimental ones; therefore, a great number of interchangeable assigned lines is obtained, and two assignment errors occur for C-10 and C-13. This result clearly demonstrates the great dependence of the capabilities of our assignment algorithm on the reference data material.

Program Output. Each assignment problem is analyzed in terms of two parameter sets. The first one describes the difference between the experimental peak list and the best calculated chemical shift for each carbon. A further parameter defines the description of the query structure by the reference data; values between 0% and 100% may occur. A value of



a) Required input-data

1) Structure: $CH_3 - CH_2 - COO - CH_2 - CH_3$
$C_1 \quad C_2 \quad C_3 \quad C_4 \quad C_5$

2) Peaklist as printed by the NMR spectrometer

L_1:	174.3 ppm
L_2:	60.1 ppm
L_3:	27.8 ppm
L_4:	14.3 ppm
L_5:	9.1 ppm

b) Assignment procedure

Step 1: Generation of the matrix, all elements are set to 1

$$\begin{array}{c|ccccc} & L_1 & L_2 & L_3 & L_4 & L_5 \\ \hline C_1 & 1 & 1 & 1 & 1 & 1 \\ C_2 & 1 & 1 & 1 & 1 & 1 \\ C_3 & 1 & 1 & 1 & 1 & 1 \\ C_4 & 1 & 1 & 1 & 1 & 1 \\ C_5 & 1 & 1 & 1 & 1 & 1 \end{array}$$

Step 2: Incorporation of multiplicity information (not used in this example, matrix remains unchanged)

Step 3: Estimation of the spectrum at the one-bond level

Estimated shift range

C_1:	0.2 - 58.8 ppm
C_2:	8.9 - 63.5 ppm
C_3:	150.1 - 191.7 ppm
C_4:	44.3 - 82.9 ppm
C_5:	0.2 - 58.8 ppm

$$\begin{array}{c|ccccc} & L_1 & L_2 & L_3 & L_4 & L_5 \\ \hline C_1 & 0 & 0 & 1 & 1 & 1 \\ C_2 & 0 & \boxed{1} & 1 & 1 & 1 \\ C_3 & 1 & 0 & 0 & 0 & 0 \\ C_4 & 0 & \boxed{1} & 0 & 0 & 0 \\ C_5 & 0 & 0 & 1 & 1 & 1 \end{array}$$

Step 4: Check of the matrix for inconsistencies; elimination, if possible

$$\begin{array}{c|ccccc} & L_1 & L_2 & L_3 & L_4 & L_5 \\ \hline C_1 & 0 & 0 & 1 & 1 & 1 \\ C_2 & 0 & \boxed{0} & 1 & 1 & 1 \\ C_3 & 1 & 0 & 0 & 0 & 0 \\ C_4 & 0 & \boxed{1} & 0 & 0 & 0 \\ C_5 & 0 & 0 & 1 & 1 & 1 \end{array}$$

Step 5: Estimation of the spectrum at the two-bond level

C_1:	0.2 - 24.8 ppm
C_2:	19.5 - 38.4 ppm
C_3:	161.5 - 191.7 ppm
C_4:	51.7 - 72.9 ppm
C_5:	5.9 - 25.6 ppm

$$\begin{array}{c|ccccc} & L_1 & L_2 & L_3 & L_4 & L_5 \\ \hline C_1 & 0 & 0 & 0 & 1 & 1 \\ C_2 & 0 & 0 & 1 & 0 & 0 \\ C_3 & 1 & 0 & 0 & 0 & 0 \\ C_4 & 0 & 1 & 0 & 0 & 0 \\ C_5 & 0 & 0 & 0 & 1 & 1 \end{array}$$

Step 6: Check of matrix for inconsistencies (none detected)

Step 7: Estimation of the spectrum at the three-bond level

C_1:	5.8 - 11.9 ppm
C_2:	22.8 - 35.1 ppm
C_3:	169.3 - 179.3 ppm
C_4:	57.4 - 65.5 ppm
C_5:	10.4 - 17.7 ppm

$$\begin{array}{c|ccccc} & L_1 & L_2 & L_3 & L_4 & L_5 \\ \hline C_1 & 0 & 0 & 0 & 0 & \boxed{1} \\ C_2 & 0 & 0 & \boxed{1} & 0 & 0 \\ C_3 & \boxed{1} & 0 & 0 & 0 & 0 \\ C_4 & 0 & \boxed{1} & 0 & 0 & 0 \\ C_5 & 0 & 0 & 0 & \boxed{1} & 0 \end{array}$$

RESULT: $CH_3 - CH_2 - COO - CH_2 - CH_3$
9.1 27.8 174.3 60.1 14.3 ppm

Figure 1. Flow chart for the assignment algorithm. Example is propionic acid ethyl ester.

Table III. Comparison of the Assignment Given in the Literature[28] and the Computer-Assisted Assignment for 17β-Hydroxy-1,4-androstadien-3-one Using Different Multiplicity Information

			computer-assisted assignment		
carbon	assignment from footnote 28	calcd shift value	SFORD	*J* modulated	without multiplicity
C-1	155.2	155.2	155.2	155.2	155.2
C-2	128.0	127.6	128.0	128.0	128.0
C-3	185.1	186.0	185.1	185.1	185.1
C-4	124.3	124.0	124.3	124.3	124.3
C-5	168.2	168.2	168.2	168.2	168.2
C-6	33.8	32.3	33.8, 32.9, 30.8	33.8, 32.9, 30.8	33.8, 32.9, 30.8
C-7	32.9	31.8	33.8, 32.9, 30.8	33.8, 32.9, 30.8	33.8, 32.9, 30.8
C-8	36.2	35.4	36.2	36.2	36.2, 37.2
C-9	53.2	52.3	53.2, 50.8	53.2, 50.8	53.2, 50.8
C-10	43.6	43.4	43.6, 43.7	43.6, 43.7	43.6, 43.7
C-11	22.8	21.0	22.8, 24.0	22.8, 24.0	22.8, 24.0
C-12	37.2	36.3	37.2	37.2	36.2, 37.2
C-13	43.7	43.4	43.6, 43.7	43.6, 43.7	43.6, 43.7
C-14	50.8	50.7	53.2, 50.8	53.2, 50.8	53.2, 50.8
C-15	24.0	23.6	22.8, 24.0	22.8, 24.0	22.8, 24.0
C-16	30.8	30.8	32.9, 30.8	32.9, 30.8	32.9, 30.8
C-17	81.2	81.1	81.2	81.2	81.2
C-18	11.4	12.0	11.4	11.4	11.4
C-19	18.8	18.7	18.8	18.8	18.8

100% means complete description of the structure at the five-bond level. The second part of the output contains the assignment, the best estimated spectrum, and the reference data used. All spectra useful during the assignment process at the three-bond level are listed with their entry number, compound name, and bibliography.

Statistical Considerations. The algorithm described above was extensively tested at different levels and works well even for complex compounds like steroids or alkaloids. The results of these computations are compiled in Table V. The data of Table V show that the number of errors decreases strongly with a growing reference data set. The number of interchangeable assigned lines does not depend so strongly on the same fact. The data also reveal the high information contents of multiplicities taken from *J*-modulated ^{13}C NMR spectra.

Erroneous assignment is mainly caused by insufficient representation of the query structure by the reference data and missing information on stereochemistry,[29,30] which is not included at the moment. A further source of errors is the chosen limits for the extension of the shift ranges according to the given formulas. Evaluation of these formulas was performed with a test data set of 250 spectra. A further extension of the shift range would decrease the number of errors but increase drastically the number of interchangeable assigned lines;

Table IV. Comparison of Assignment Given in the Literature[31] and the Computer-Assisted Assignment for Roquefortine[a]

carbon	assignment from footnote 31	calcd shift value	computer-assisted assignment
C-1	114.5	113.1	114.5
C-2	143.2	144.9	143.2
C-3	40.9	43.3	40.9
C-4	22.9	19.3	22.9, 22.5
C-5	22.5	19.3	22.9, 22.5
C-6	61.5	54.1	61.5
C-7	36.8	31.9	36.8
C-8	58.8	65.0	58.8
C-9	166.7	168.0	166.7, 159.2
C-10	121.9	142.2	149.8, 128.5
C-11	110.9	123.4	110.9, 134.3, 136.4, 109.1, 128.9
C-12	125.5	125.0	121.9, 125.5, 149.8, 128.5
C-13	134.3	119.9	110.9, 128.9
C-14	136.4	137.4	134.3, 136.4, 128.9
C-15	159.2	162.0	166.7, 159.2, 149.8
C-16	78.3	87.1	78.3
C-17	149.8	137.9	121.9, 125.5, 159.2, 149.8, 128.5
C-18	109.1	113.2	110.9, 134.3, 136.4, 109.1, 128.9
C-19	128.9	127.0	134.3, 136.4, 128.9
C-20	119.0	119.7	119.0
C-21	125.0	123.2	125.0
C-22	128.5	137.6	125.5, 149.8, 128.5

[a] Carbons are numbered according to structure input. Multiplicity information is from SFORD. All values are in ppm.

Table V. Result of Assignment Procedure Using Different Numbers of Reference Data and Different Multiplicity Informations

no. of ref spectra	multiplicity	assignment errors (%)	assigned lines per carbon
2039	SFORD	2.14 (1:47)	1.35
4234	SFORD	1.83 (1:55)	1.29
6313	SFORD	1.39 (1:72)	1.30
6313	*J* modulated	1.47 (1:68)	1.30
6313	without	1.47 (1:68)	1.48

therefore, these formulas are a well-working compromise.

CONCLUSIONS

The program CSEARCH has proven its excellent flexibility during a 3-year period. The various search strategies available allow subtle feature selection. The most powerful tool of this software package is the fully automated assignment procedure for carbon-13 nuclear magnetic resonance spectra, which is now routinely used in our research group as well as by other scientists. This algorithm performs the same strategy as the spectroscopist, starting with multiplicity information and then including shift arguments at different levels. The program does not release the chemist from checking the results; however, differences between the experimental values and the calculated ones can help in planning further experiments to decide ambiguity.

EXPERIMENTAL

The CSEARCH program consists of about 180 subroutines and is written in FORTRAN-IV. The routines for file handling and bit manipulation are programmed in ASSEMBLER. The program runs on a CDC CYBER-170/720 computer under NOS at the University Computing Center. It requires about 33K central memory words (60 bits per word).

ACKNOWLEDGMENT

We express our gratitude to Dr. E. Haslinger (University of Vienna) and Dr. U. Jordis and Dr. H. Fröhlich (Technical University of Vienna) for helpful discussions. We thank Dr. C. A. Shelley (Eastman Kodak Company, Rochester, NY) for a copy of his structure display program and Dr. N. Müller (University of Linz) for some routines performing data manipulation. We are indebted to C. Vecera for compiling ^{13}C NMR data. This work was supported by the Fonds zur Förderung der wissenschaftlichen Forschung in Österreich, Project 5388.

REFERENCES AND NOTES

(1) Gray, N. A. B. "Computer-Assisted Analysis of Carbon-13 NMR Spectral Data". *Prog. Nucl. Magn. Reson. Spectrosc.* **1982**, *15*, 201–248.

(2) Bremser, W. "Automatische Aufnahme und Interpretation von NMR-Spektren". *Chem.-Ztg.* **1980**, *104*, 53–61.

(3) Bremser, W. "The Importance of Multiplicities and Substructures for the Evaluation of Relevant Spectral Similiarities for Computer Aided Interpretation of ^{13}C-NMR Spectra". *Z. Anal. Chem.* **1977**, *286*, 1–13.

(4) Dubois, E.; Bonnet, J. C. "The DARC Pluridata System: The ^{13}C-NMR Data Bank". *Anal. Chim. Acta* **1979**, *112*, 245–252.

(5) Zippel, M.; Mowitz, J.; Köhler, I.; Opferkuch, H. J. "SPEKTREN—A Computer System for the Identification and Structure Elucidation of Organic Compounds". *Anal. Chim. Acta* **1982**, *140*, 123–142.

(6) Finer-Moore, J.; Mody, N. V.; Pelletier, S. W.; Gray, N. A. B.; Crandell, C. W.; Smith, D. H. "Computer-Assisted Carbon-13 Nuclear Magnetic Resonance Spectrum Analysis and Structure Prediction for the C_{19}-Diterpenoid Alkaloids". *J. Org. Chem.* **1981**, *46*, 3399–3406.

(7) Gray, N. A. B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. "Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data". *J. Org. Chem.* **1981**, *46*, 703–715.

(8) Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Crandell, C. W. "The DENDRAL Project: Recent Advances in Computer-Assisted Structure Elucidation". *Anal. Chim. Acta* **1981**, *133*, 471–497.

(9) Milne, G. W. A.; Heller, S. R. "NIH/EPA Chemical Information System". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204–211.

(10) Milne, G. W. A.; Zupan, J.; Heller, S. R.; Miller, J. A. "Spectra-Structure Relationships in Carbon-13 Nuclear Magnetic Resonance Spectroscopy. Results from a Large Data Base". *Org. Magn. Reson.* **1979**, *12*, 289–296.

(11) Heller, S. R.; Milne, G. W. A. "The NIH-EPA Chemical Information System in Support of Structure Elucidation". *Anal. Chim. Acta* **1980**, *122*, 117–138.

(12) Zupan, J.; Heller, S. R.; Milne, G. W. A.; Miller, J. A. "A Substructure-Oriented ^{13}C-NMR Chemical Shift Retrieval System". *Anal. Chim. Acta* **1978**, *103*, 141–149.

(13) Shelley, C. A.; Munk, M. E. "Computer Prediction of Substructures from Carbon-13 Nuclear Magnetic Resonance Spectra". *Anal. Chem.* **1982**, *54*, 516–521.

(14) Gribov, L. A.; Elyashberg, M. E.; Koldashov, V. N.; Pletnjov, I. V. "A Dialogue Computer Program System for Structure Recognition of Complex Molecules by Spectroscopic Methods". *Anal. Chim. Acta* **1983**, *148*, 159–170.

(15) Abe, H.; Fujiwara, I.; Nishimura, T.; Okuyama, T.; Kida, T.; Sasaki, S. "Recent Advances in the Structure Elucidation System CHEMICS". *Comput. Enhanced Spectrosc.* **1983**, *1*, 55–62.

(16) Neszmelyi, A.; Kmety, A. "On-Line Data Retrieval for ^{13}C-NMR Spectroscopy". Private Communication.

(17) Bremser, W.; Wagner, H.; Franke, B. "Fast Searching for Identical ^{13}C-NMR Spectra via Inverted Files". *Org. Magn. Reson.* **1981**, *15*, 178–187.

(18) Functional group: a carbon and its surrounding atoms.

(19) Bremser, W. "HOSE—A Novel Substructure Code". *Anal. Chim. Acta* **1978**, *103*, 355–365.

(20) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS-Online Search System. 1. General System Design and Selection, Generation and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(21) Attias, R. "DARC Substructure Search System: A New Approach to Chemical Information". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

(22) Hagadone, T. R.; Howe, W. J. "Molecular Substructure Searching: Minicomputer-Based Query Execution". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 182–186.

(23) Howe, W. J.; Hagadone, T. R. "Molecular Substructure Searching: Computer Graphics and Query Entry Methodology". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 8–15.

(24) Shelley, C. A. "Heuristic Approach for Displaying Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 61–65.

(25) Benn, R.; Günther, H. "Moderne Pulsfolgen in der hochauflösenden NMR-Spektroskopie". *Angew. Chem.* **1983**, *95*, 381–411.

(26) Robien, W. "Computerunterstützte Zuordnung von ^{13}C-NMR Spektren". *Monatsh. Chem.* **1983**, *114*, 365–372.

(27) Lindley, M. R.; Gray, N. A. B.; Smith, D. H.; Djerassi, C. "Applications of Artificial Intelligence for Chemical Inference. 40. Computerized Approach to the Verification of Carbon-13 Nuclear Magnetic Resonance Spectral Assignments". *J. Org. Chem.* **1982**, *47*, 1027–1035.

(28) Hickey, J. P.; Butler, I. S.; Pouskouleli, G. "Carbon-13 NMR Spectra of Some Representative Hormonal Steroids". *J. Magn. Reson.* **1980**, *38*, 501-506.

(29) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. "Stereochemical Substructure Codes for ^{13}C Spectral Analysis". *Org. Magn. Reson.* **1981**, *15*, 375-389.

(30) Beierbeck, H. "Simple Stereochemical Structure Code for Organic Chemistry". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 215-222.

(31) Broadbent, T. A.; Paul, E. G. "Carbon-13 Nuclear Magnetic Resonance in Alkaloid Chemistry". *Heterocycles* **1983**, *20*, 863-980.

A Concise Connection Table Based on Systematic Nomenclatural Terms

J. D. RAYNER

Department of Computer Studies, The University of Hull, Hull HU6 7RX, U.K.

Received July 6, 1984

A connection table schema is introduced that can be used to describe concisely both molecular structures and mixtures, organic and potentially also inorganic, within a single uniform hierarchy. The schema can explicitly represent heteroatoms, charges, and different bond types but is nonexplicit in its coding of the most common individual carbon atoms in the molecule, being based on systematic nomenclatural terms for substructural units such as rings and chains.

BACKGROUND

During the course of an investigation[1] into the computer translation of IUPAC systematic nomenclature,[2,3] a target representation of molecular structure was required that could be formed as output from the developing nomenclature translation system. The immediate need was not for a molecular structural representation that could be fully explicit at the atomic level but rather a scheme that could include information at the level of detail found in the systematic name.

IUPAC names are formed from small fragments that often relate individually to polyatomic structural units—ring systems, carbon chains, etc.—in which the nature of bonding and the atom types are assumed to be generally uniform. Other fragments deal with deviations from these assumptions or denote smaller structural units such as individual atoms. Thus, in the concise connection table (CCT) described below, the bulk of the carbon skeleton in an organic structure is represented implicity, as it is in the nomenclature, but all non-carbon atoms and all unusual bonds (that is multiple or nonaromatic according to context) and all electronic charges that are described by a systematic name can be represented explicity in the table.

In contrast, other connection table schemes are generally fully explicit in their description of every (non-hydrogen) atom in a structure, and some contain redundancy in their duplication of bond information.[4] Such a redundant connection table may list every non-hydrogen atom in a structure, and by means of index numerals and bond multiplicity values, it can represent all the interatom connections. However, since every recorded atom has all its associated bonding represented within its own record, information on any one bond is duplicated, in the records of the two atoms that it joins. Such redundancy can be eliminated by recording only one bond in each atom record, for example, that to the adjacent atom with lowest index, but further records of a different type are then necessary to specify ring closure in cyclic compounds.

One advantage of the CCT described below is its implicit retention of ring closure information, while avoiding the use of any variant records. Further compactness, beyond the avoidance of such variants, is achieved in the CCT schema by dispensing with the bulk of bond information entirely, on the basis that for organic compounds (and therefore the bulk of those actively dealt with) the great majority of bonding is entirely predictable from the nature and configuration of the atoms themselves, in their four common environments—aromatic, aliphatic, alicyclic, and individual. The relatively few variations from these default states may still be represented with a net reduction in table size, as for instance in examples 8 and 9 of Figure 3 below.

GENERAL DESCRIPTION

A concise connection table (CCT) will comprise a sequence of table entries that are each of identical form but whose meaning will vary according to both content and context within the abstract hierarchy of the table. Since the CCT schema has been developed for use as the output of a nomenclature translation process, it is not unnatural for the table structure to be hierarchic, in resemblance of the approach of the IUPAC nomenclature. However, this is not to say that the IUPAC nomenclature is a necessary precursor to the use of the CCT schema for representing any molecular structure: any system that deals in substructural units may be potentially applicable, for example, the coding of chain fragments and ring system skeletons in the Wiswesser line notation (WLN).[5]

In the present application, the nomenclatural term that represents the parent structure of a molecule gives rise to the initial entries of the corresponding table. Substituents on the parent are represented by further entries, which are both hierarchically inferior and physically subordinate in the table. The relative ordering of individual table entries is further discussed below. The terms "parent" and "substituent" derive from a consideration of organic structures, but the CCT is also potentially capable of representing inorganic compounds.

The contractions possible for organic chains and rings of carbon cannot generally be echoed in the coding of the more varied assemblages of atoms and bonds that characterize inorganics. Bonding arrangements in the latter may be more accurately represented by further codings in the schema beyond those listed below, given appropriate extensions to the present definition.

The greater variety of structure found in an inorganic substance may thus cause a proliferation of table entries, with the result that a CCT for an inorganic substance can closely approximate to a full atom-by-atom representation. Nevertheless, since inorganic "parental" structures are typically single atoms, repeated substituents (ligands etc.) will generally have the unit locant, and their representation can then be condensed by means of the special entry for repetition (*qv*). Assemblages

0095-2338/85/1625-0108$01.50/0 © 1985 American Chemical Society