Automated structure elucidation system – CHEMICS

Kimito Funatsu, Carlos A. Del Carpio, and Shin-ichi Sasaki

Laboratory for Chemical Information Science, Toyohashi University of Technology, Toyohashi 440, Japan

Automatisiertes System zur Strukturaufklärung – CHEMICS

Zusammenfassung. Ein rechnerunterstütztes Programm zur Strukturaufklärung organischer Substanzen, CHEMICS, wurde von den Autoren entwickelt. Es besteht im wesentlichen aus den Teilen Datenanalyse und Strukturgenerierung. Im vorliegenden Beitrag wird die gegenwärtige Form von CHEMICS und ein neuer Ansatz für die Auswertung von NMR-Daten (^{1}H und ^{13}C) unbekannter Verbindungen beschrieben, wobei der Schwerpunkt bei der Datenanalyse liegt. Dieses Programm kann sinnvolle Partialstrukturen, in einigen Fällen auch Gesamtstrukturen aufzeigen sowie chemische Verschiebungen sinnvoll zuordnen. Hierbei werden charakteristische Relationen zwischen den chemischen Verschiebungen für ^{1}H und ^{13}C spezifizierte Partialstrukturen und eine neuentwickelte Korrelationstabelle verwendet. Durch die Erweiterung von CHEMICS um diesen Modul wird die Auswahl aus den Gesamtstrukturen, die sich aus den Teilstrukturen konstruieren lassen, stark konzentriert.

Summary. A computer-assisted structure elucidation system of organic compounds, CHEMICS, has been developed. The system mainly consists of two parts, i.e., data analysis and structure generation. The paper describes the outline of current CHEMICS and a new approach for the analyses of unknown's NMR data (^{1}H and ^{13}C), focussing on the role of data analysis. The program for the new approach can elucidate reasonable partial (in some case, full) structure(s) together with the reasonable assignments of chemical shifts, using characteristic relationships between both NMR chemical shifts for specified substructures and chemical shift-substructure index files newly prepared for this approach. By introduction of this module to CHEMICS, the number of candidate structures constructed based on these partial structures will show a considerable decrease.

Introduction

Many studies have been done for structure elucidation and identification of organic compounds with aid of computers. The methodologies and the techniques are classified into two categories. One is the retrieval method in which the identification is carried out by refining the most likely structure from a data base by comparing data, for instance, chemical spectra, of an unknown with those of organic compounds stored there. The other is a structure generation method; that is, the most probable structure is generated by the automated analysis of data (also, for instance, chemical spectra) of an unknown using empirical and theoretical rules.

The CHEMICS system, developed and published by the authors [1], is a computer-assisted structure elucidation system for organic compounds, which depends mainly on the latter way. The principle of the system is that all possible structures, which are known to exist or which might exist on chemical grounds, are listed in a computer. The number of structures in a particular case is then narrowed down by successively entering information from spectroscopic measurements.

CHEMICS is designed to store all the substructures (called 'components') necessary for building any likely structures. At present, CHEMICS contains 630 components for the structure elucidation of organic compounds consisting of C, H, O, N, S and halogen atoms (Table 1). The set of components has been devised so that it is possible to construct any structures by selecting appropriate components from the complete set. To store such a set of components in a computer is synonymous with storing all the complete structures which could be present.

Outline of current CHEMICS

The current CHEMICS system is composed of the following four functional modules, as shown in Fig. 1: *a*) Data analysis, *b*) Structure generator, *c*) Stereo-generator, *d*) Input of macrocomponent (partial structure).

Data analysis

Among the components which have survived because they are consistent with the molecular formula, some can be subsequently discharged because they are inconsistent with ^{1}H- and ^{13}C-NMR chemical shift values, or IR data. In the selection of components by CHEMICS these spectral data measured on the sample are compared by computer with those in component/chemical shift or component/wave number correlation tables, so that only components consistent with these data are left. Part of the correlation table showing ppm ranges for ^{1}H and ^{13}C shifts is shown in Table 2.

Offprint requests to: S. Sasaki

The next step is to make component sets by use of the components which have been selected as being not contradictory to the molecular formula and spectral data. Since such a principle of component selection discards only the components that are contradictory to data, some of those components which remain appear to be useless to a chemist's eyes. However, this situation is inevitable because it is the object of this system never to miss the correct solution. Then, in the first step towards making the component sets, all combinations of all components which can make up the molecular formula are examined. CHEMICS does not collect all the random combinations but involves a logic to exclude prohibited combinations when co-ordination-prohibited component pairs are found in allocating each of the components to the corresponding NMR signals in view of NMR data [10].

Table 1. Component set for structure elucidation of organic compounds containing C, H, O, N, S, and halogens

No.	Component		No.	Component	
1	tert-Bu−	(S)	372		(I)
2		(ND)	373	$>C-$	(Br)
⋮			374		(Cl)
51	CH_3CH_2-	(CD)	⋮		
52		(CT)	403	$>N \rightarrow O$	(Y)
53		(CS)			
⋮			404	$>S \rightarrow O$	(Y)
185	$CH_3-\underset{\underset{O}{\downarrow}}{\overset{\overset{O}{\uparrow}}{S}}-$	(O)			
186		(Y)	⋮		
			547	−OH	(CD)
			548		(CT)
⋮			549		(CS)
351	$>C=NH$	(F)	⋮		
352		(S)	626	−F	
353		(ND)	627	−Cl	
⋮			628	−Br	
			629	−I	
			630	−D−	

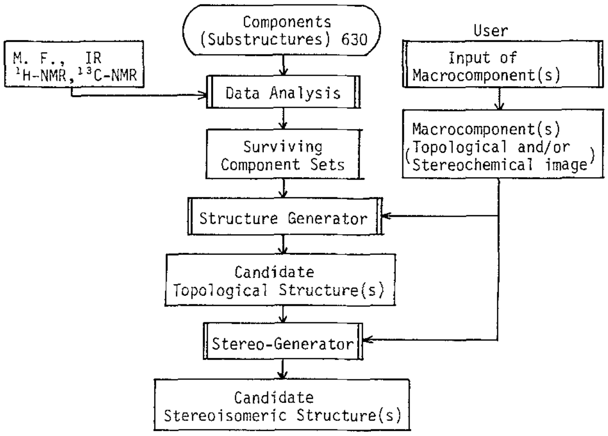


Fig. 1. Block diagram of the current CHEMICS

Structure generator

The next step is to generate structures from the individual component combinations. The generation is carried out taking all possibilities into account, in due consideration of the principle that most of the components can only be linked to a limited number of species. On the basis of a specially designed logic, connectivity stack [5], when the system functions properly it does not reproduce the same structure nor does it fail to generate any structure which can justifiably be built.

Stereo-generator

The major role of the above module is generation of constitutional isomers (topological image). On the other hand, this module has a function for generating all possible stereoisomeric structures due to asymmetric carbon, double bond and so on using topological information of the respective constitutional isomers generated by 'structure generator'. The detailed algorithm has been reported in previous paper [2].

Although improbable formulae may be involved in final solutions from 'structure generator' or 'stereo-generator', we should treat them all as having exactly the same probability at this stage, so long as they are all consistent with the molecular formula and spectral data of the sample through the current analytical procedures.

Input of macrocomponent (partial structure) [7]

The chemist often has some information about the structure of a sample. This may be obtained from the past record of the sample or the experience in its laboratory handling. When the partial structure is entered by the user, the constitutional information is degraded into its components as shown in Table 1, which are then compared with the

Table 2. Correlation table for NMR analyses

No.	Components		^{1}H-NMR (ppm)				^{13}C-NMR (ppm)	
16	$(CH_3)_2CH-$	(CS)	1.20	0.50	28.9	13.4	40.7	18.2
132	CH_3CO-	(O)	2.50	1.80	24.2	17.7	174.0	165.8
196	$>CH_2$	(N)	5.60	1.10	75.5	25.7		
197	$>CH_2$	(O)	6.10	2.30	88.6	43.2		
198	$>CH_2$	(Y)	5.40	1.70	60.4	6.6		
199	$>CH_2$	(CD)	6.20	0.50	58.0	11.9		
208	$-CH<$	(N)	7.30	1.10	96.9	27.1		
209	$-CH<$	(O)	7.70	1.80	111.1	41.3		
211	$-CH<$	(CD)	4.40	0.80	75.8	15.4		
274	$-CH=$	(N)	9.60	6.60	184.5	91.6		
277	$-CH=$	(CD)	9.00	4.50	165.0	90.1		

components that the system has selected. The system will adopt the information entered only when all the components derived from the partial structure inserted have already been selected by CHEMICS. This means that the components which the system has selected with a full safety factor will take precedence over the additional information which has been entered manually. The stereochemical information of the macrocomponent is reflected on the final results according to the other logic[1].

As obvious from the above explanation, the fragments for making up structures and the carriers of spectral information are just components. The unit for examining the reasonable allocation of each component to NMR signals, is also component centering around the correlation table. Moreover, the examination of the input macrocomponent by spectral data is also based on the component unit. It is obvious that essentially the analytical ability of 'data analysis' in CHEMICS never exceeds what is provided by component units. Thus, correspondence of candidate structures with input data is said to become ambiguous in some cases. The number of candidates increases in proportion to the ambiguity. According to the principle of never missing correct solution, this result is said to be unavoidable. As one of the ideas for coping with this situation, CHEMICS-F [6], which has a file retrieval function, has been developed, and the modules for prediction of the number of ^{13}C-NMR signals [4][2] and judgement of probability on the basis of strain energy calculation [1], although off-line with the present CHEMICS version, have been provided. These functions play an effective role after generation of whole structures.

On the other hand, the introduction of partial structures selected by the user, has enhanced the correctness and practically of structure elucidation by our system. However, if possible, it seems to be one of the ideal features that partial structures entered should be determined by agreement with both deduction by the computer and judgement of it by the user. In order to realize about this situation an analytical way different from that in 'data analysis' of the current CHEMICS is required. In this sense, as a new approach of automated partial structure elucidation, an interdependent analytical way based on the relationships between ^{1}H- and ^{13}C-NMR chemical shifts for each atomic group with specified neighboring groups, has been developed.

Automated partial structure elucidation on the basis of interdependent analysis of ^{1}H- and ^{13}C-NMR spectra

It is widely known that today ^{1}H- and ^{13}C-NMR spectral data are inevitable to structure elucidation. Lots of papers have reported trials of computer-assisted structure elucidation with the two kinds of NMR data. Most of the trials, however, handle separately the ^{1}H- and ^{13}C-NMR data analysis. Accordingly, structural information to be obtained through mutually relating one NMR data to the other has not been brought into conventional methods.

The authors have found out the correlation between the corresponding ^{1}H- and ^{13}C-NMR chemical shifts exhibited by a substructure over many compounds and used the correlation for computer-assisted structure elucidation in order to analyze and relate the two spectra complimentarily to each other.

Table 3. Atoms and atomic groups used in this module[a]

Code No.		Code No.		Code No.	
1	>C<	17	=N−	33	$-NH_2$
2	>CH−	18	=NO−	34	−NO
3	>C=	19	−ACH−	35	$-NO_2$
4	>N−	20	−AO−	36	−NS
5	>NO−	21	−AS−	37	−NSO
6	>AC−	22	−AN	38	$-N_3$
7	−CH=	23	−ANH−	39	$=CH_2$
8	−C≡	24	−ANO−	40	=C=O
9	$-CH_2-$	25	−ASO−	41	=NH
10	>C=O	26	$-ASO_2-$	42	=N=N
11	−O−	27	CH_3-	43	=S
12	−S−	28	−CH=O	44	≡CH
13	−NH−	29	−C≡N	45	−F
14	−SO−	30	−N≡C	46	−Cl
15	$-SO_2-$	31	−OH	47	−Br
16	=C=	32	−SH	48	−I

[a] Symbol A in codes means aromatic

Correlation between ^{1}H- and ^{13}C-NMR chemical shifts

The ^{1}H- and ^{13}C-NMR chemical shifts denote the magnetic environment around ^{1}H and ^{13}C nuclei, respectively. Accordingly, on the assumption that the magnetic reflection of structural environment on the target atomic groups can be approximately treated by the atomic group units set in Table 3, there must be a correlation between carbon and hydrogen chemical shifts contained in eight atomic groups, CH_3-, $-CH_2-$, −CH<, −CH=, $=CH_2$, ≡CH, −CHO, −ACH− (symbol A in code means aromatic).

Figure 2 shows an illustration of the corresponding ^{1}H- and ^{13}C-NMR chemical shifts for each atomic group over about 300 kinds of compounds for which both NMR were measured. Expectedly, there exists, though coarse but noticeable, a relationship between the two chemical shifts. This fact suggests that specification of structural environment may give unique features relating the structural environment with the corresponding ^{1}H- and ^{13}C-NMR chemical shifts. Therefore, this relationships has been examined by cluster analysis [9] on the basis of magnetic assortment of the neighbors for each of the target atomic groups[3]. Figure 3 shows the analytical result for only the target atomic group of methyl, and the numbers show the cluster number to which the molecules belong. The correlation functions between ^{1}H- and ^{13}C-NMR chemical shifts within each of clusters are calculated by Newton interpolation polynomial method. As obvious from Fig. 3, it can be recognized that each correlation function is proper to the environment of the corresponding substructures. In other words, both NMR chemical shifts corresponding to a substructure within a compound must satisfy the characteristic correlation of the substructure. Inversely, it can be said that the fact is effective

[1] To be published

[2] At present, the number of ^{13}C-NMR signals can be predicted under consideration of stereochemistry, corresponding to the output level of final candidates

[3] The detailed analytical procedures and the application method of the results to automated spectral analysis will be reported in another paper

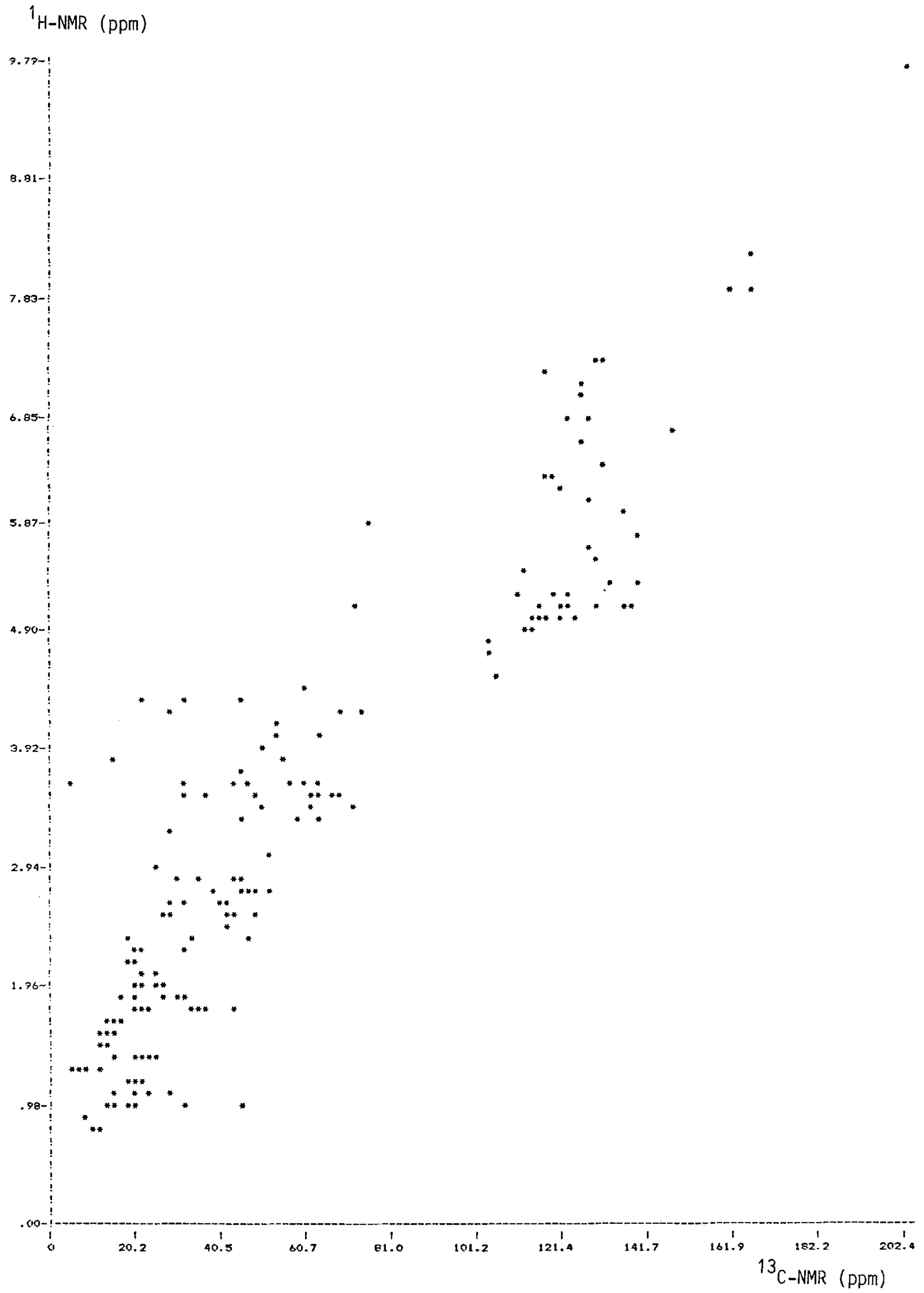


Fig. 2
A relationship between ^{1}H- and ^{13}C-NMR chemical shifts (because of the over-position of many points due to the scale one point represents many chemical shift values of similar molecules)

to identify the structural environment meeting with the ^{1}H- and ^{13}C-NMR chemical shifts of an atomic group contained in an unknown compound. We will abbreviate the ^{1}H-NMR-^{13}C-NMR chemical shift correlation to "P-C correlation" here in this investigation and describe the method for identifying a partial structure by the use of the P-C correlation.

Block diagram of automated partial structure elucidation

The block diagram is shown in Fig. 4. When the molecular formula, ^{1}H-NMR chemical shifts and ^{13}C-NMR chemical shifts and their multiplicities of unknown compound are input, the system generates all the possible candidate atomic group sets which are necessary to make up structures, by the comparison of atomic groups (Table 3) previously stored in a computer with those data[1]. The system carries out the P-C correlation analysis of the set, referring to the ^{1}H- and ^{13}C-NMR chemical shift-substructure index files (Fig. 5) and obtains a partial structure. Finally, the system outputs the assumedly most appropriate partial structure together with the assignments of chemical shifts.

Atomic group units

The authors have built a ^{1}H- and ^{13}C-NMR data base, out of which the chemical shift-substructure index file is edited.

[1] This routine utilizes part of the program reported in [8]

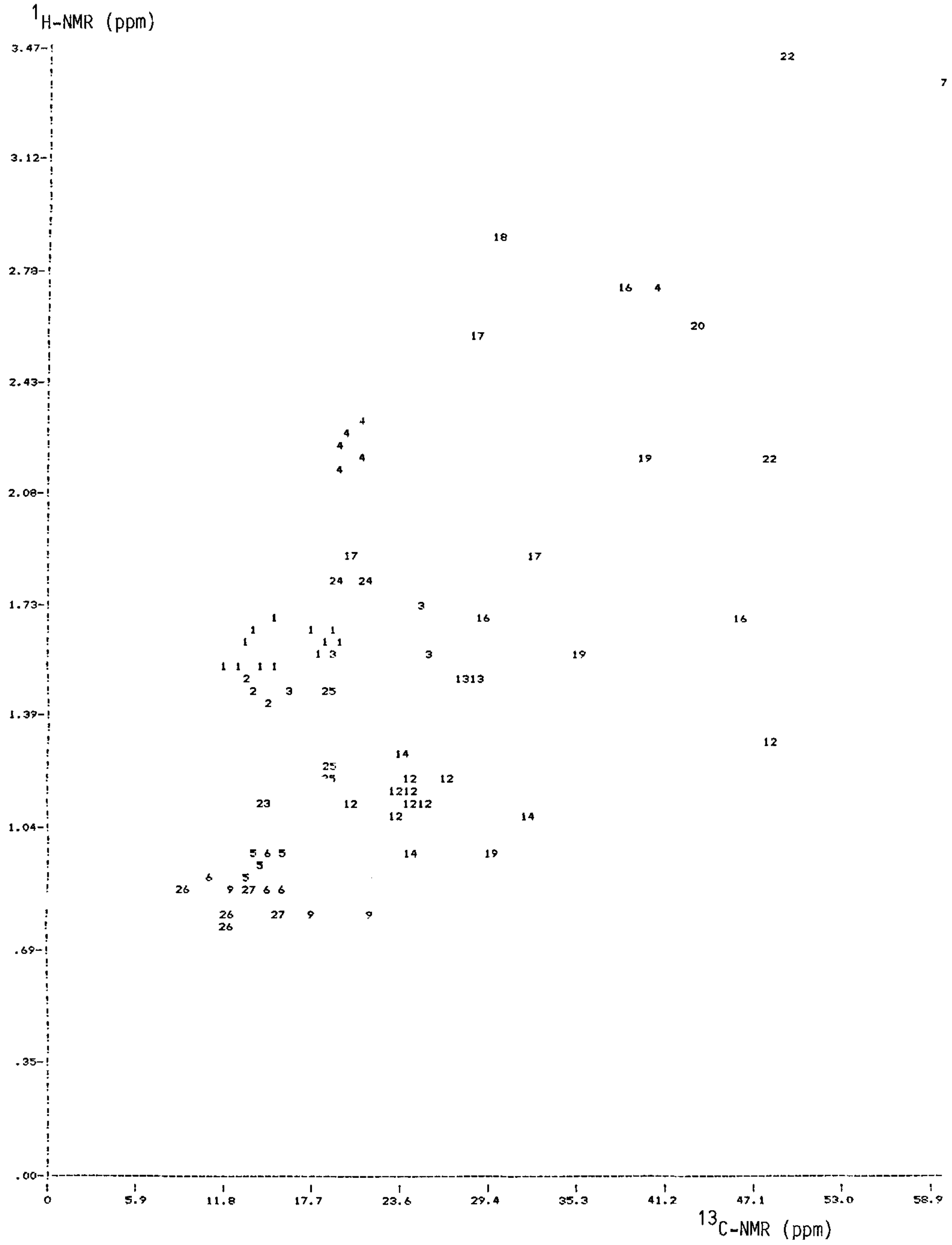


Fig. 3
The result of cluster analysis for the target atomic group of methyl (because of the overposition of many points due to the scale one point represents many chemical shift values of similar molecules. The members show the cluster number to which the molecules belong)

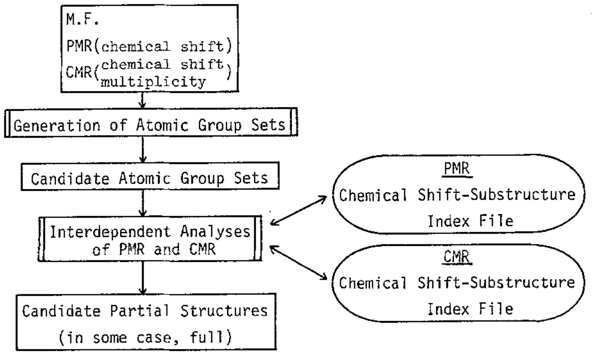
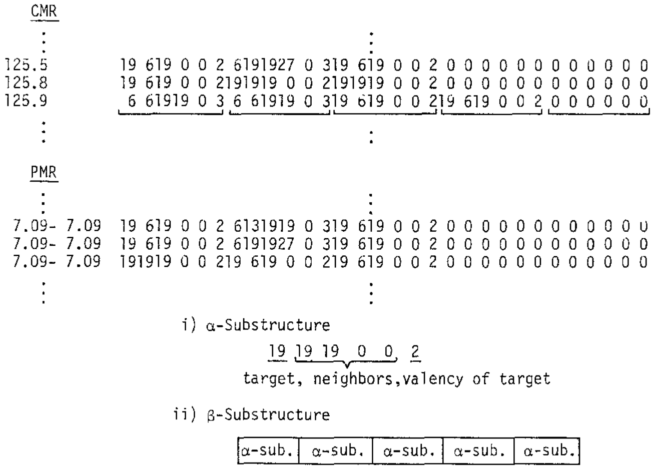

M.F.
PMR(chemical shift)
CMR(chemical shift multiplicity)
Generation of Atomic Group Sets
Candidate Atomic Group Sets
Interdependent Analyses of PMR and CMR
Candidate Partial Structures (in some case, full)
PMR Chemical Shift-Substructure Index File
CMR Chemical Shift-Substructure Index File

Fig. 4. File structure of index files

In addition, atomic group units handled in the module are settled in compliance with the CANOST code [3] (Table 3).

4. Chemical shift-substructure index files

The ^1H- and ^{13}C-NMR index files are prepared on the basis of appropriate 3,000 data, respectively. Figure 5 shows parts of the two NMR index files, in which substructures, each expressed by an atomic group code in Table 3, the chemical shifts, and the atomic groups in the α-, and β-positions are described. Each substructure is described as a string of a two-digit atomic group code (Table 3) composed of five blocks. In this paper, each of the blocks and full strings are called α- and β-substructures as shown at the foot of Fig. 5. The authors will explain the file structure, taking up strings



```
CMR
  :                                   :
125.5        19 619 0 0 2 6191927 0 319 619 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
125.8        19 619 0 0 2191919 0 0 2191919 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
125.9         6 61919 0 3 6 61919 0 319 619 0 0 219 619 0 0 2 0 0 0 0 0 0
  :                                   :
PMR
  :                                   :
7.09- 7.09   19 619 0 0 2 6131919 0 319 619 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
7.09- 7.09   19 619 0 0 2 6191927 0 319 619 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
7.09- 7.09   191919 0 0 219 619 0 0 219 619 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
  :                                   :

                 i) α-Substructure
                        19 19 19  0  0  2
                 target, neighbors,valency of target
                 ii) β-Substructure
                      α-sub. | α-sub. | α-sub. | α-sub. | α-sub.
```

Fig. 5. Input data and first α-substructures

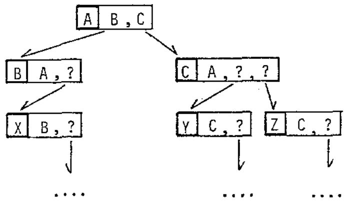
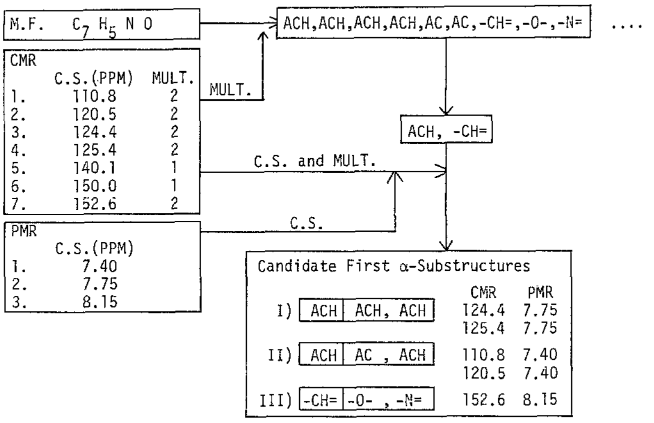


Fig. 6. Conceptual chart of search of next substructures

shown in Fig. 5, line 1, as an example. The target atomic group, 19(ACH) exhibits a chemical at 125.5 ppm. Position α of the target 19 contains two substructures 6 (AC) and 19 (ACH) (the first block), which are indicated by a string of 619, and a figure of 2 in the lowest column refers to the valency of the target 19. Substructure 6 in position α has bonds with substructures 19, 19, and 27 (CH_3) (the second block). One of the two substructures 19s is the target 19 itself. In other words, two substructures 27 and the other 19 are in position β when viewed from the target 19. A figure of 3 denotes the valency of substructure 6. In addition, substructure 19 in position α has bonds with substructures 6 and 19 (the third block). The latter substructure 19 is the target 19 itself. Accordingly, only substructure 6 is in position β when viewed from the target 19. A figure of 2 is the valency of substructure 19. This file structure is very effective because it allows the search of the environment by unit of blocks in view of bonds. The file structure also allows circumstance denotation for extension.

Procedure of partial structure generation by the use of P-C correlation

The system generates an atomic group set, which is necessary to make up a structure, according to the molecular formula, ^{13}C-NMR multiplicities, and bond conditions of the unknown compound. The system basically takes the following method for generation of partial structure from an atomic group set (Fig. 6): if position α of an assumed target atomic group (A) is specified like B or C, the system connects α-substructures (Fig. 5) into a tree shape, thereby generating a partial structure. Accordingly, an α-substructure (called the first α-substructure) to be the root of the tree needs being specified first. Then, the system selects atomic groups containing C and H atoms together out of the atomic group set already generated, and submits the atomic groups to bond examination (α) and P-C correlation examination (α).



CMR

	C.S.(PPM)	MULT.
1.	110.8	2
2.	120.5	2
3.	124.4	2
4.	125.4	2
5.	140.1	1
6.	150.0	1
7.	152.6	2

PMR

	C.S.(PPM)
1.	7.40
2.	7.75
3.	8.15

Candidate First α-Substructures

			CMR	PMR
I)	ACH	ACH, ACH	124.4	7.75
			125.4	7.75
II)	ACH	AC , ACH	110.8	7.40
			120.5	7.40
III)	-CH=	-O- , -N=	152.6	8.15

Fig. 7. Block diagram of this module

Bond examination (α). This routine checks target atomic groups for validity of an α-substituent through looking into the bond conditions. The most important items are: (i) Checking for bond condition in an atomic group set on the basis of the kinds of atomic group and the number of their bonds.

(ii) Checking for β-substituent settlement conditions in view of α-substituent settlement (Fig. 6).

P-C Correlation examination (α). The procedure for the examination of the validity of an α-substituent through the application of P-C correlation is:

(*i*) The routine selects a ^{13}C-NMR chemical shift, which contains the same ^{13}C-NMR multiplicity as the designated target atomic group has, out of the data on sample compound. Next, the routine inquires the ^{13}C-NMR index file of the selected chemical shift, and searches for an α-substructure which has the designated atomic group and has passed bond examination (α). If searching is unsuccessful, control is returned to the start of this stage and the routine carries out the next chemical shift selection.

(*ii*) This routine calculates the ^{1}H-NMR chemical shift, which corresponds to the ^{13}C-NMR chemical shift of the picked up α-substructure, by the P-C correlation method. If there is the calculated ^{1}H-NMR chemical shift in the sample ^{1}H-NMRs, control is passed to (*iii*) and, if not, control is returned to (*i*) to search for another ^{13}C-NMR chemical shift.

(*iii*) This routine inquires the ^{1}H-NMR index file of whether the chemical shift calculated from the P-C correlation corresponds to the α-substructure or not. If the chemical shift corresponds, the routine assigns the α-substructure to the ^{13}C- and ^{1}H-NMR chemical shifts and presumes the α-substructure to exist in the unknown compound. If not, the routine presumes not and returns control to (*i*) again to take the same steps for another α-substructure.

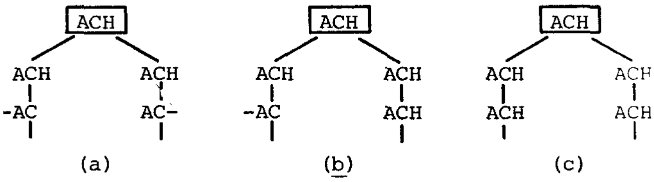
Figure 7 shows the above procedure with a compound having a molecular formula of C_7H_5NO as an example. [ACH,ACH,ACH,ACH,AC,AC, – CH = , – O – , – N=] is one of the candidates of an atomic group set, generated mainly from ^{13}C-NMR multiplicities. The routine selects two atomic groups, ACH and – CH=, which contain C and H atoms together and are to be the roots of a tree and submits the candidates to bond examination (α) and P-C correlation examination (α). As indicated in *I*), *II*) and *III*), this routine outputs three kinds of first α-substructure and the chemical shifts to be assigned to them. The upper part of Fig. 8 shows a flow diagram of this procedure.

Selection of next α-substructures for each branch-"ALPHA". This routine goes on structure elucidation with the first α-substructure (*I*). The two α-substituents in (*I*) are ACHs; that is, two branches hang from it. Each branch becomes a target atomic group (T′) of the α-substructures to be connected next, and one of the α-substituents of the next α-substructure must be designated as a target atomic group (T) in the base (*I*) (bond examination (α) – (*ii*)). The remaining substituent is determined from spectral data (P-C correlation examination (α)). As a result, the next candidate α-substructures connecting to each branch of the first α-substructure (I) (α-substructures connected to the first branches) and candidates for chemical shifts to be assigned to the substructures are determined as listed below:

Parent α-substructure No. 1		ACH/ACH, ACH	
		^{13}C-NMR (ppm)	^{1}H-NMR (ppm)
Branch 1	ACH/AC, ACH	110.8	7.40
		120.5	7.40
	ACH/ACH, ACH	124.2	7.75
		125.4	7.75
Branch 2	ACH/AC, ACH	110.8	7.40
		120.5	7.40
	ACH/ACH, ACH	124.2	7.75
		125.4	7.75

The parent α-substructure refers to the α-substructure which contains the target atomic group (T′) of the α-substructures connected to the first branches; that is the first α-substructure (I).

By the use of the list mentioned above, the environment from the target atomic group (T) to position β of the parent α-substructure can be expressed as three β-substructures (a) to (c) shown below.



Note: Rectangles each denote the target atomic group of the parent α-substructure No. 1

Calculation of β-substructures-"BETA". Figure 8 shows in its center "BETA" which carries out β-substructure calculation. First, the routine carries out the bond examination (β), which corresponds to bond examination (α), of each of β-substructures (a) to (c) for the validity of β-substituents and γ-subsituents. If the β-substructures (a) to (c) pass bond examination (β), the routine carries out the P-C correlation examination (α). In this case, these three β-substructures passed bond examination (β). If the β-substructures pass the two examinations, the routine presumes the substructure to exist in the unknown compound. In the illustrated case, only β-substructure (b) is found to exist in the unknown compound. Consequently, the routine selects ACH/AC,ACH and ACH/ACH,ACH as candidates for the second α-substructures connecting to the parent α-substructure No. 1 and the candidates as the tips for the next growth of the tree structure. The ^{1}H- and ^{13}C-NMR chemical shifts which conform to the P-C correlation function supporting the existence of the β-substructure are assigned to the target atomic group:

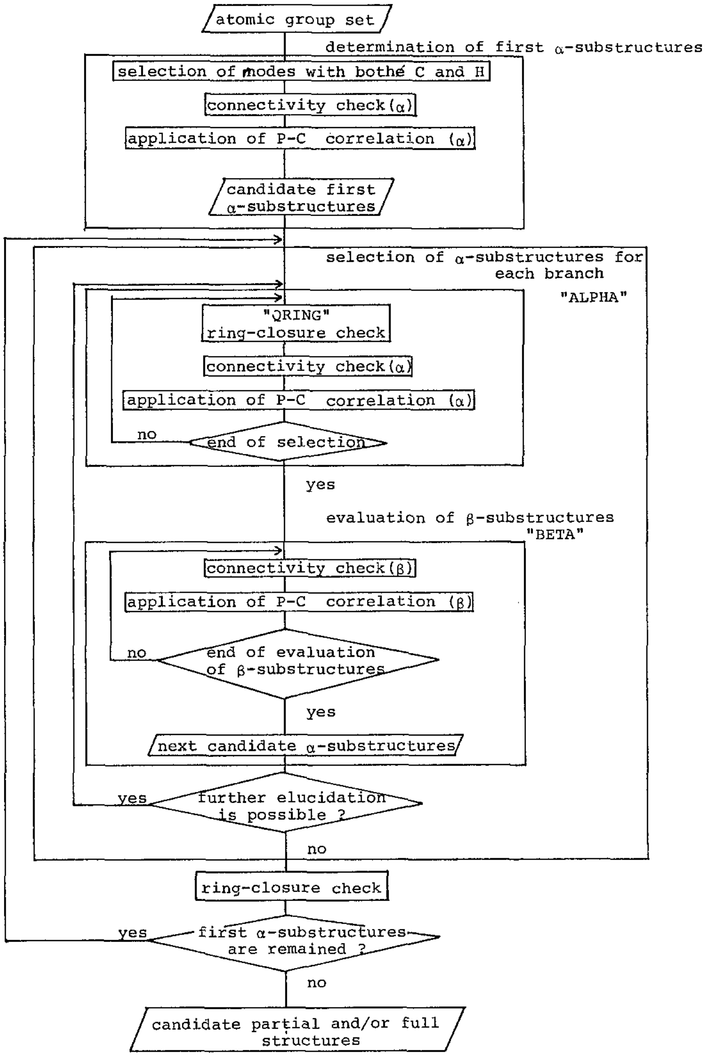

atomic group set
determination of first α-substructures
selection of modes with bothé C and H
connectivity check(α)
application of P-C correlation (α)
candidate first α-substructures
selection of α-substructures for each branch
"ALPHA"
"QRING" ring-closure check
connectivity check(α)
application of P-C correlation (α)
no
end of selection
yes
evaluation of β-substructures
"BETA"
connectivity check(β)
application of P-C correlation (β)
no
end of evaluation of β-substructures
yes
next candidate α-substructures
yes
further elucidation is possible ?
no
ring-closure check
yes
first α-substructures are remained ?
no
candidate partial and/or full structures

Fig. 8. Flow chart of independent analysis

Parent α-substructure No. 1	ACH/ACH, ACH	
	^{13}C-NMR (ppm)	^{1}H-NMR (ppm)
Chemical shift assignment	124.2	7.75
	125.4	7.75
Next candidate α-substructures		
No. 2 ACH/AC, ACH		
No. 3 ACH/ACH, ACH		

Now two kinds of chemical shift have been assigned to the target atomic group of the parent α-substructure; however, the step is inevitable in the present status as far as the routine uses the current P-C correlation functions. In addition, if two or more different β-substructures pass β-substructures examinations, the routine cannot discriminate the correct one from others for the time being. Accordingly, the routine adopts only the α-substructures which are common to all successful β-substructures, for the next growth of the structure tree and prunes the others because of indefinite courses. When the routine has pruned every branch else, the routine completes the structure elucidation.

If the tree structure keeps on growing, the system repeats the basic procedure. The system, lets the check routine (Fig. 8), QRING, check β-substructures for the possibility of ring-closure formation, that is a bond between branches, for each step and does not repeat the procedure carelessly.

Ring-closure check routine-"QRING". The routine sometimes uses, repeatedly as an α-substituent of the current parent α-substructure, a specific atomic group which has been used in the stage of structure growth. On that occasion, if there is a free bond on the atomic group, the routine carries out a ring-closure operation after bond examination (β) and P-C correlation examination (β). If the ring-closure operation is unsuccessful, the routine rejects the request and stops the growth of the branches. Ring-closure formation is basically carried out by the spanning tree method. Since there is discretion in judging a ring-closure formation from molecular size and functionality, a variety of troubles arise.

Thus far, the authors have described an atomic group set containing, as the targets, atomic groups with C and H atoms together. From now on, the authors will cover an atomic group set of which the atomic groups, as the targets, do not contain both C and H atoms or at all.

Elucidation of the environment of atomic groups consisting of heterogeneous atoms. When heterogeneous atomic groups are adopted as the targets, naturally the routine cannot use ^{1}H- and ^{13}C-NMR data. It is possible, however, for the routine to list up candidates for the α-substructure by designating the already found neighboring α-substructure targets as α-substituents and by restricting the number and kinds of remaining α-substituents on the basis of the results of bond examination (α). If the α-substructure determined from the ^{1}H- and ^{13}C-NMR data adjoins the target of the β-substructure although the target is a heterogeneous atomic group, it is possible for the routine to infer the existence of the heterogeneous atomic group from that fact together with the results of bond examination (β). If there exists no β-substructure to presume, the routine prunes the branches which seem to be irrational at the next structure reasoning step. If two or more heterogeneous atomic groups reasonably grow as bonds, information from NMR data cannot specify the structure of bonds, and the structure inferred is less dependable. In those cases, the adoption of IR spectral data is suggested.

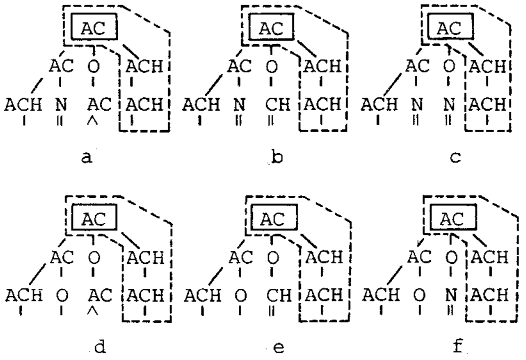
Elucidation of the environment of atomic groups containing either of H and C atoms. The routine calculates by bond examinations (α) and (β) and from either ^{1}H or ^{13}C-NMR data, the environment of atomic groups containing either of H and C atoms.

Since some constituents of α- and β-substituents are already determined as in the preceding subsection, however, it is not so difficult to carry out environment examination up to the position β.

In the illustrated case, the routine settles an oxygen atom (−O−) as an α-substituent adjoining an α-substructure No. 4 which further adjoins a parent α-substructure No. 2. When the routine regards the α-substructure No. 4 as a new parent α-substructure, the routine lists up possible branch α-substructures as shown below:

New parent α-substructure No.4		AC/AC, −O−, ACH	
		^{13}C-NMR (ppm)	^{1}H-NMR (ppm)
Branch 1	AC/AC, =N−, ACH	140.1	–
	AC/AC, −O−, ACH	150.0	–
Branch 2	−O−/AC, AC	–	–
	−O−/AC, =CH−	–	–
	−O−/AC, =N−		

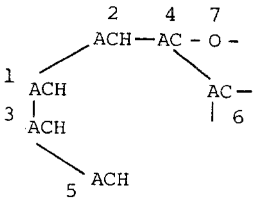
The routine carries out the examination (β) of β-substructures (a) to (f), which are made up on the basis of the above step, refers to the ^{13}C-NMR index file, and judges the β-substructure (b) to exist. In addition, the routine assigns a chemical shift to the parent α-substructure No. 4 and determines other α-substructures No. 6 and 7 which are to be connected additionally to the parent α-substructure No. 4.



Notes: (1) *Solid rectangles* denote the target atomic group of the parent α-substructure No. 4. (2) *Dotted rectangles* denote α-substructures determined already

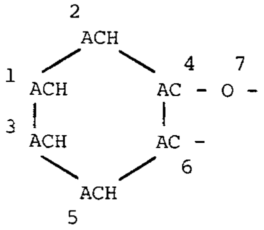
Parent α-substructure No. 4	AC/AC, −O−, ACH	
	^{13}C-NMR (ppm)	^{1}H-NMR (ppm)
Chemical shift assignment	150.0	–
Next candidate α-substructures		
No. 6 AC/AC, =N−, ACH		
No. 7 −O−/AC, =CH−		

Shown below are the results inferred up to this stage for a compound with a molecular formula of C_7H_5NO.

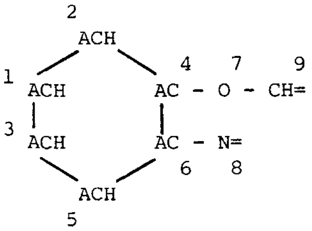


Note: Figures written alongside denote the order of atomic groups generated

In the target 5 environment inference stage, the "QRING" routine suggests a ring closure existing between atomic groups 5 and 6. The routine judges the ring closure to be rational through carrying out the procedure in the above subsection and forms the following structure:



When the last α-substructure inference has been completed, the opposite parties of all bonds are specified. If the bonds can be formed according to the specification, the routine completes structure inference. If not, the routine examines the possibility of forming a ring closure. In the illustrated case, since connection is not complete, as shown below, the routine requires forming a ring closure. Shown below is a structure before the formation of a ring structure:



The routine picks up the position of ring-closure occurrence and submits the β-substructures, which are to adjoin the atoms at those positions, to bond examination (β) and P-C correlation examination (β). If the β-substructures pass the examinations, the routine outputs the final results. If not, the routine shows to the operator the results in the stage preceding to ring-closure formation. In the illustrated case, a ring closure is formed between atomic groups 8 and 9.

Figure 9 shows a tree graph explaining the accepted and rejected processes, for α-substructures in the structure inference stated above, together with those for first α-substructures (II) and (III). Figure 10 shows together with the assignments of chemical shifts the final results inferred from the first α-substructures (I) to (III). The whole structure is obtained from (I) and (III), and the assigned chemical shifts agrees with the actual values. In addition, the partial structure inferred from (II) constitutes a part of the whole structure obtained from (I) and (III), which supports the validity of the inferred whole structure.

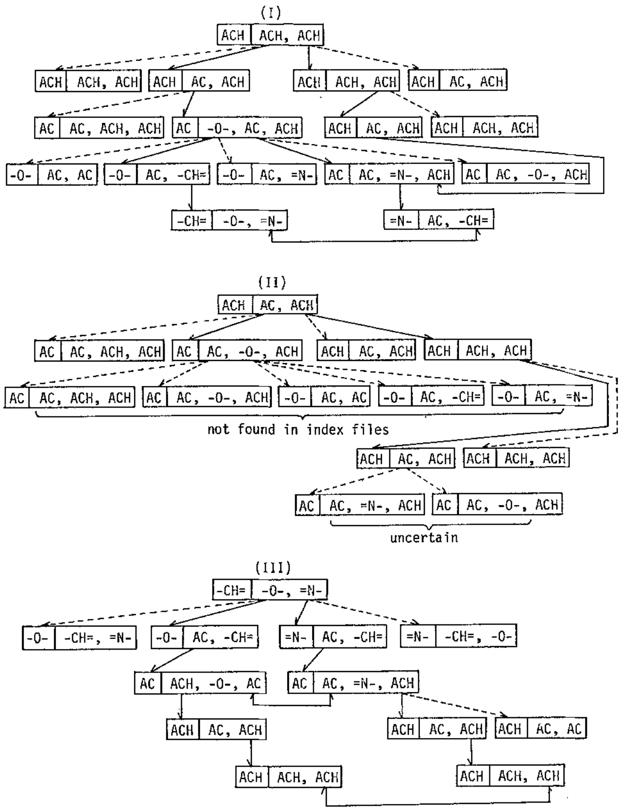
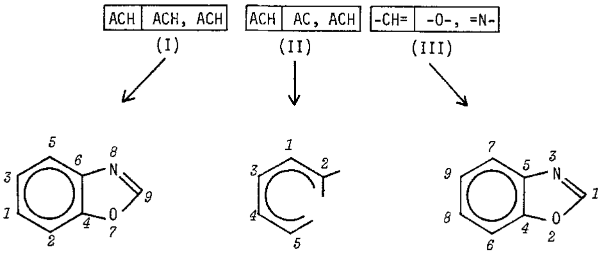


Fig. 9. Tree graphs of structure elucidation



	CMR	PMR
1	124.2 125.4	7.75 7.75
2	110.8	7.40
3	124.2 125.4	7.75 7.75
4	150.0	—
5	120.5	7.40
6	140.1	—
7	—	—
8	—	—
9	152.6	8.15

	CMR	PMR
1	110.8	7.40
2	—	—
3	124.2	7.75
4	125.4	7.75
5	110.8 120.5	7.40 7.40

	CMR	PMR
1	152.6	8.15
2	—	—
3	—	—
4	150.0	—
5	140.1	—
6	110.8	7.40
7	120.5	7.40
8	124.4	7.75
9	125.5	7.75

Fig. 10. Final results of illustrated case

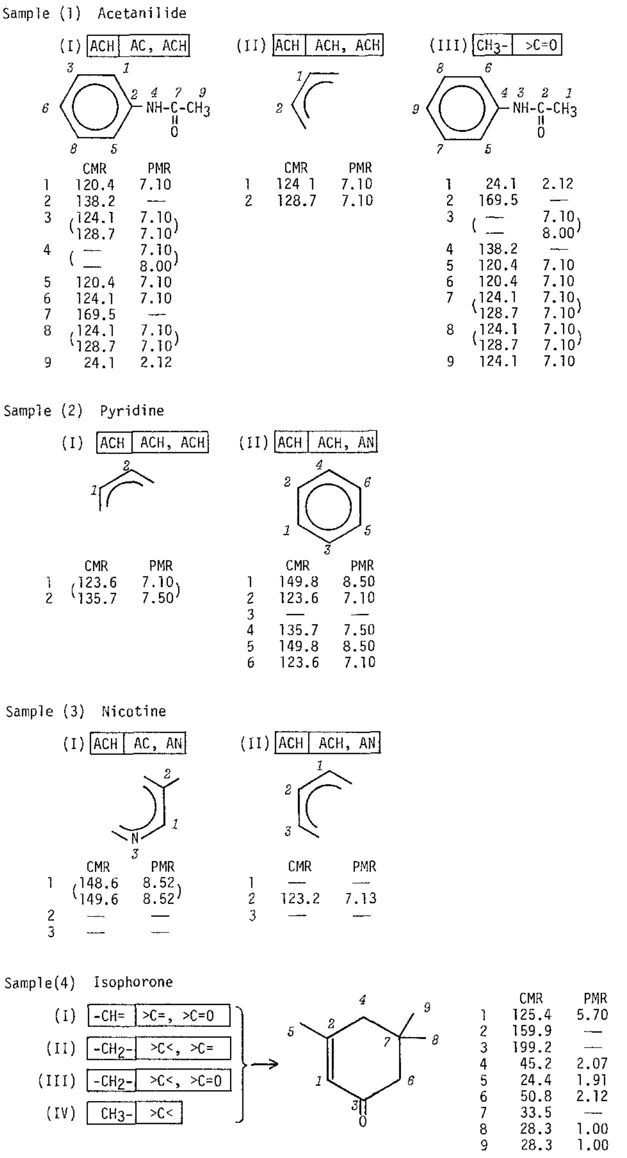
Experimental examples

The authors tried operating the system with the following four samples. Figure 11 shows the respective results: (*1*) acetanilide, (*2*) pyridine, (*3*) nicotine, and (*4*) isophorone.

For example (*3*), the system listed up eight first α-substructures, two of which were successful in forcasting the existence of the substituted pyridine ring. For example (*4*), the system inferred the whole structure independently from four first α-substructures. The figure shows the chemical shift assignment obtained only from the results of (*1*). In all examples, the assignments of chemical shifts agrees with actual measurements.

Problems

The method described so far is an excellent means for elucidating a partial structure from NMR spectra. P-C correlation functions are determined from actual measurements. When the amount of actual measurement data is higher, the quality of the correlation functions increases and the discrimination power of the system rises. As of today,



Sample (1) Acetanilide (I)

	CMR	PMR
1	120.4	7.10
2	138.2	—
3	(124.1 128.7	7.10 7.10)
4	(— —	7.10 8.00)
5	120.4	7.10
6	124.1	7.10
7	169.5	—
8	(124.1 128.7	7.10 7.10)
9	24.1	2.12

Sample (1) Acetanilide (II)

	CMR	PMR
1	124 1	7.10
2	128.7	7.10

Sample (1) Acetanilide (III)

	CMR	PMR
1	24.1	2.12
2	169.5	—
3	(— —	7.10 8.00)
4	138.2	—
5	120.4	7.10
6	120.4	7.10
7	(124.1 128.7	7.10 7.10)
8	(124.1 128.7	7.10 7.10)
9	124.1	7.10

Sample (2) Pyridine (I)

	CMR	PMR
1 2	(123.6 135.7	7.10 7.50)

Sample (2) Pyridine (II)

	CMR	PMR
1	149.8	8.50
2	123.6	7.10
3	—	—
4	135.7	7.50
5	149.8	8.50
6	123.6	7.10

Sample (3) Nicotine (I)

	CMR	PMR
1	(148.6 149.6	8.52 8.52)
2	—	—
3	—	—

Sample (3) Nicotine (II)

	CMR	PMR
1	—	—
2	123.2	7.13
3	—	—

Sample(4) Isophorone

	CMR	PMR
1	125.4	5.70
2	159.9	—
3	199.2	—
4	45.2	2.07
5	24.4	1.91
6	50.8	2.12
7	33.5	—
8	28.3	1.00
9	28.3	1.00

Fig. 11. Experimental examples

the system only carries out structure examination from NMR data as far as β-position. For this purpose, the system uses the chemical shift-substructure index files as a means for solving the problem.

The next subject still to take up is to use NMR information more; for instance, the meaning of signal intensity needs intent using. The authors plan to introduce a program for the purpose in the near future.

The recognition of cyclic structure of polycyclic compounds is not easy as far as the tree-type inference system is used. Generation of duplicative environment is not avoidable in some case. The problem still remains to be solved.

Conclusion

To try analyzing the two kinds of spectrum in view of the correlation between ^{1}H- and ^{13}C-NMR chemical shifts has been successful through utilizing the intentionally edited chemical shift-substructure index files.

The present approach has aimed at carrying out spectral and connectivity checks against relatively large substructures (β- and γ-substructures) and obtaining the partial structure information which assist to exclude the impractical outputs by CHEMICS. Not only partial structure but chemical shift assignment to these seems to be useful as additive information in evaluation of the whole structure.

The size of elucidated partial structure depends on the accuracy of P-C correlation functions of the index files. Therefore, the size may be changeable with the refinement of them. Although some problems remain to be solved, it is made possible to allow the computer to elucidate the partial structures which are up to now selected only by the user as outer information and evaluate them by the user. This module will be edited into the author's automated structure elucidation system, 'CHEMICS' in the near future.

References

1. Abe H, Fujiwara I, Nishimura T, Okuyama T, Kida T, Sasaki S (1983) Comput Enhanced Spectrosc 1:55
2. Abe H, Hayasaka H, Miyashita Y, Sasaki S (1984) J Chem Inf Comput Sci 24:216
3. Abe H, Kudo Y, Yamasaki T, Tanaka K, Sasaki M, Sasaki S (1984) J Chem Inf Comput Sci 24:212
4. Fujiwara I, Okuyama T, Yamasaki T, Abe H, Sasaki S (1981) Anal Chim Acta 133:527
5. Kudo Y, Yamasaki T, Sasaki S (1973) J Chem Doc 13:225; Kudo Y, Sasaki S (1974) J Chem Doc 14:200; Kudo Y, Sasaki S (1976) J Chem Inf Comput Sci 16:43
6. Sasaki S, Abe H, Hirota Y, Ishida Y, Kudo Y, Ochiai S, Saito K, Yamasaki T (1978) J Chem Inf Comput Sci 18:211
7. Sasaki S, Fujiwara I, Abe H, Yamasaki T (1980) Anal Chim Acta 122:87
8. Abe H, Okuyama T, Fujiwara I, Sasaki S (1984) J Chem Inf Comput Sci 24:220
9. Varmuza K (1980) Pattern recognition in chemistry (Lecture notes in chemistry, No. 21). Springer, Berlin Heidelberg New York
10. Yamasaki T, Abe H, Kudo Y, Sasaki S (1977) ACS symposium series no. 54. Computer-assisted structure elucidation: 108

Received December 2, 1985