

Application of the automated structure elucidation system (CHEMICS) to the chemistry of natural products

Kimito Funatsu, Yutaka Susuta, and Shin-ichi Sasaki

Toyohashi University of Technology, Tempaku, Toyohashi 440, JAPAN

Abstract

Principles of the automated structure elucidation system for organic compounds (CHEMICS) and its applications to natural product chemistry are described. It is made clear that CHEMICS is at the practical level in structure elucidation of natural products through the introduction of 2D-NMR data processing and specific partial structure elucidation functions to the system.

The CHEMICS system, developed by the authors, is a computer-assisted structure elucidation system for organic compounds, which depends on the way of 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.¹⁾ 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 the 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.

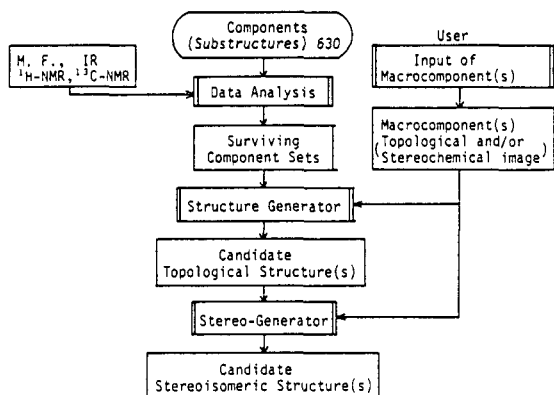


Fig.1. Block diagram of the current CHEMICS.

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	(Br)
...		374	(Cl)
51	CH ₃ CH ₂ - (CD)	...	
52	(CT)	403	N→O (Y)
53	(CS)	404	S→O (Y)
...		...	
185	O (O)	...	
186	↑ (Y)	547	-OH (CD)
	CH ₃ -S-	548	(CT)
	↓	549	(CS)
	O	...	
...		...	
351	C=NH (F)	626	-F
352	(S)	627	-Cl
353	(ND)	628	-Br
...		629	-I
		630	-D-

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).

a) 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 H-1 and C-13 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 H-1 and C-13 shifts is shown in Table 2. 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.

Table 2. Correlation table for NMR analyses.

No.	Components		¹ H-NMR (ppm)		¹³ C-NMR (ppm)		
16	(CH ₃) ₂ CH-	(CS)	1.20	0.50	28.9	13.4	18.2
132	CH ₃ CO-	(O)	2.50	1.80	24.2	17.7	174.0
196	>CH ₂	(N)	5.60	1.10	75.5	25.7	165.8
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	

b) Structure generator: This 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 specially designed logic, connectivity stack, 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.

c) Stereo-generator: The major role of the above module is generation of constitutional isomers. 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'.

d) Input of macrocomponent (partial structure): 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 (described above), which are then compared with the 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 other logic.

As obvious from the above explanation, the fragments for making up structures and carriers of spectral information are just components. The unit of 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, which has a file retrieval function, has been developed, and the modules for prediction of the number of C-13 NMR signals and judgement of probability on the basis of strain energy calculation, 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 practicality 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 H-1 and C-13 NMR chemical shifts for each atomic group with specified neighboring groups, has been developed. According to this method, rather big-sized partial structure can be generated automatically which is most helpful for narrowing the number of candidates. Furthermore, the introduction of carbon-carbon signal connectivity information provided by 2-D NMR into CHEMICS has been accomplished so that the connectivity of carbons in skeletal structure of an unknown is automatically analyzed and elucidated at the both steps of 'data analysis' and 'structure generator' in generating whole candidate structures. This also results in the fewer number of candidates.

EXAMPLES OF STRUCTURE ELUCIDATION

The first example of the elucidation process of an unknown compound (actually, nicotine) is serially illustrated in Fig. 2. Chemical data of the unknown, molecular formula and spectroscopic data of IR, H-1 and C-13 NMR, are compared with 630 components at the step of 'data analysis' to give rise 65 components. The number of candidates will amount to more than ten thousand if all the possible structures are constructed on the basis of the 65 components. In this example, the interdependent analytical method based on the relationships between H-1 and C-13 NMR chemical shifts to suggest automatically the presence of '3-substituted pyridyl group' as a possible substructure in the unknown structure so that the candidates' number drastically diminishes into only seven structures in which the underlined correct solution is involved.

The second example deals with an unknown compound a part of which is already identified from chemical or instrumental analysis made by chemist. The remaining part is generated automatically and connected to the known part to form the whole structure. Let's assume an unknown compound which has the molecular formula $C_{17}H_{17}O_6Cl$ and comprises a known substructure, denoted as A, as the host (Fig. 3). If the composition of A, i.e. $C_9H_7O_4Cl$, and the number of

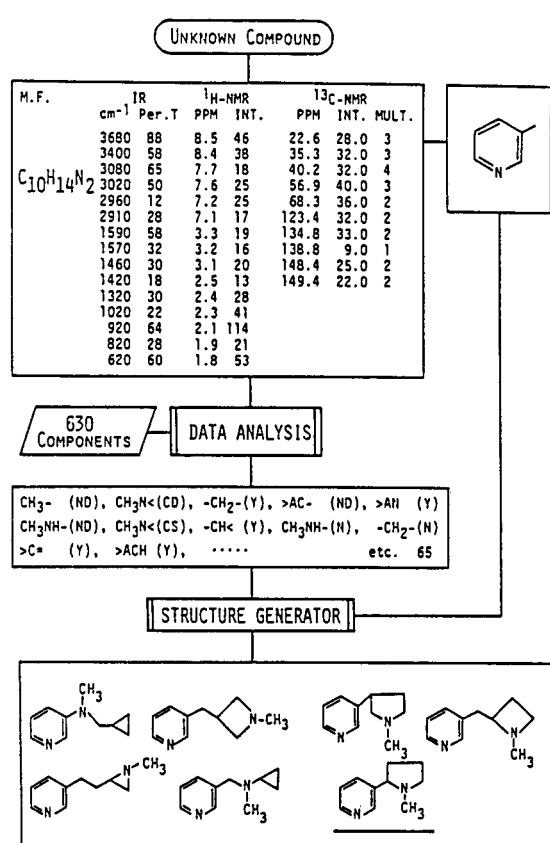


Fig. 2. The analytical processes using 3-substituted pyridyl group provided by the H-1 and C-13 NMR chemical shift interdependent analysis.

being left for construction of the counterpart that is expressed as $C_8H_{10}O_2$. Thus, the operation, which is performed with reference to NMR data, generates four substructures (B, C, D, E) as candidates for the counterpart, as shown in Fig. 3. Finally, the four are connected with the host (A) to generate six structures, F, G, H, J, K and L, of which G correctly represents the structure of griseofulvin used as the unknown compound, as listed in Fig. 4.

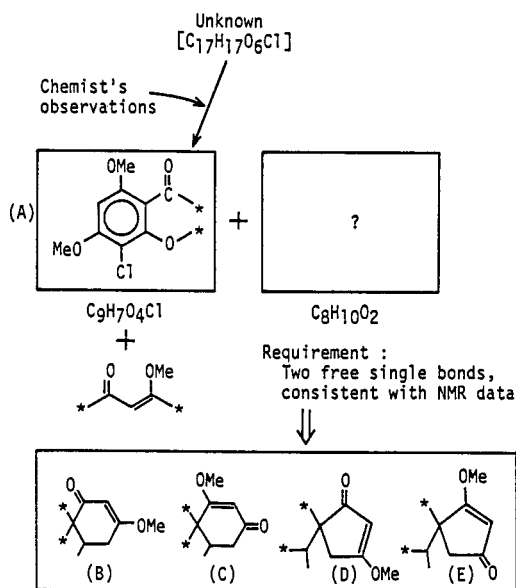


Fig.3. The conditions and results for elucidating the counterpart structure.

its single bonds, i.e. two, are entered in the computer, calculation will be performed on the assumption that the remaining part (counterpart) has the composition of $C_8H_{10}O_2$ and two single bonds. Analysis of this compound by CHEMICS in terms of its molecular formula, H-1 NMR and C-13 NMR allows 65 components to remain as the components consistent with these data. However, of the 65, 16 components are not consistent with input data in consideration of forming the counterpart, with the remaining 49 components

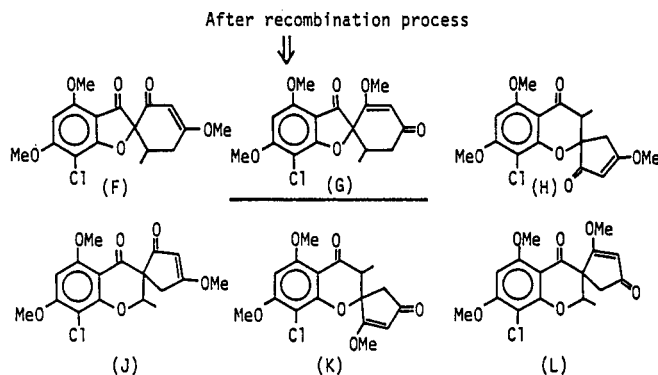


Fig.4. The candidates provided by connection of counterpart structures with host (A).

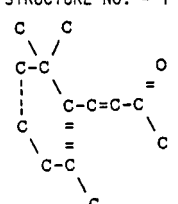
The third example shows a procedure for applying 2-D NMR data.²⁾ CHEMICS analysis of H-1 and C-13 NMR spectra of a certain compound (C₁₃H₂₀O) gives 39 components, from which 1417 candidate structures are generated. 2-D NMR data can serve to reduce this number drastically. Three types of carbon-carbon signal connectivity information (on C-13 NMR) to be entered in the computer are obtained from the 2-D NMR of this specimen as shown below.

- Type 1: 1-10, 4-11, 7-9 (long range C-H cosy : path 2)
 Type 2: 2-5, 6-12, 8-12 (combination of C-H and H-H cosy)
 Type 3: 3-4, 3-7, 4-8, 6-7 (long range C-H cosy : path 3)

Table 3. The results of 2-D NMR analyses using H-H cosy and C-H cosy information.

Information used in CHEMICS	The number of candidates
Ordinary Analyses	1,417
1 - 10 4 - 11 7 - 9	203
2 - 5 6 - 12 8 - 12	36
3 - 4 4 - 8 3 - 7 6 - 7	1

STRUCTURE NO. = 1

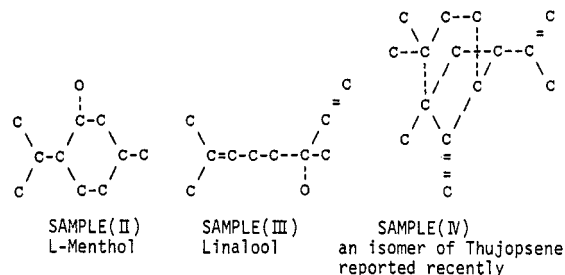


SAMPLE(I) : β-IONONE

Receiving these data the computer carries out the following operations. The data of Type 1 are used mainly to decide on the possibility of existence of each component with two carbons contained in the group of 39 components. As a result, from the remaining 36 components 203 structures are generated without using Type 2 and Type 3 data. Next, if the Type 2 and Type 3 data are used together with Type 1 data, the number of generated structures is decreased to 36 by examination through the Type 2 data and then to only one, the target compound beta-ionone(I), through Type 3 data. the above results are summarized in Table 3.

Table 4. The results of 2-D NMR analyses using 2D-INADEQUATE information.

Information used in CHEMICS	The number of candidates		
	II	III	IV
Ordinary Analyses	219	48	4,450
by Check 1	46	3	920
by Check 1 Check 2	5	3	17
by Check 1 Check 2 Check 3	3	2	9
by Check 1 Check 2 Check 3 Check 4	1	1	1



In Table 4 is summarized the number of candidate structures given by CHEMICS for menthol (II), linalool(III) and an isomer of thujopsene(IV), reported recently, with and without the aid of 2D-INADEQUATE information. In these samples, the number of candidates diminishes into one and only correct structure through four check stages.

REFERENCES

- 1) K. Funatsu, N. Miyabayashi, and S. Sasaki, *J.Chem.Inf.Comput.Sci.*, **28**, 19(1988).
- 2) K. Funatsu, Y. Susuta, and S. Sasaki, *J.Chem.Inf.Comput.Sci.*, submitted.