The Computer-assisted Characterization of Terpenes and Related Compounds by the Use of Combined Spectrometric Data

Shin-ichi SASAKI,* Hidetsugu ABE,*** Keiji SAITO,** and Yoshiaki ISHIDA**
Miyagi University of Education, Aoba, Sendai 980

(Received April 10, 1978)

Basic research to prepare a combined retrieval system based on three different sets of spectrometric data, infrared, mass, and ^{1}H NMR, for terpenes and related compounds is described. The system may be used either for searching all three types of spectra, for two, or for one only. The results of 112 test searches are presented. The combined method is proved to be a good tool for the identification of such compounds, especially those in a structurally close relationship.

A number of search systems for the structure characterization of chemical compounds have been reported, but most of them are designed for a single type of spectra.[1—29] However, to improve the possibility of identification, several kinds of combined retrieval systems based on different spectrometric data have also been developed.[30—37] The present authors are doing basic research to make a new searching system combining three different sets of spectrometric data (NMR, IR, and MS) to characterize organic compounds as a link in the study of the CHEMICS system (a computer program system for the structural elucidation of organic compounds).[38] In this paper the identification of terpenes and the related compounds will be described. The spectrometric data of these compounds, which number over, 110, were taken from a spectral atlas;[39] all of the data were packed in a computer in the form of normalized data. The normalization of the spectrometric data was carried out in a manner to be described later. The sorting experiment was implemented by the successive comparison of one of the stored compounds with all of them. The results given by the combined comparison are shown to be completely reliable.

Data Base. Numerals, 1, 2, 3,···, and 112 were given for each terpene and related compound, as is shown in Table 1. Their NMR, IR, and mass spectra were normalized and recorded in a computer. The normalization of spectra was performed as follows: the NMR was converted into G and S values according to Eqs. 1 and 2, where G stands for the center of gravity of a whole spectrum, and S, the standard deviation of each signal to the center:[40]

$$G = \sum_{i=1}^{n} (W_i A_i) / \sum_{i=1}^{n} A_i, \tag{1}$$

$$S = (\sum_{i=1}^{n} (W_i - G)^2 A_i / \sum_{i=1}^{n} A_i)^{1/2}. \tag{2}$$

Here, A_i and W_i express the weight of the proton(s) and the position of the ith signal, respectively. For example, the NMR of β-selinene (101)(Fig. 1) is converted into $G(=1.778)$ and S $(=1.121)$.

The IR is normalized by expressing it with the positions and intensities of the absorption bands. The spectral region of 4000 to 650 cm^{-1} is divided into 18 main divisions, at 3200, 2800, 2300, 2000, 1900, 1800, 1700, 1600, 1500, 1400, 1300, 1200, 1100, 1000, 900, 800, and 700 cm^{-1}. Further, each division is divided into 10 equal parts to make subdivisions, 1—10. The position (P) of the strongest band in each division is expressed numerically with the number of the subdivision where the top of the band is located. To code the band intensity (I), the following equations (Eqs. 3 and 4) are necessary to convert the apparent transmittance (T) into A' *via* the pesudo absorbance (A):

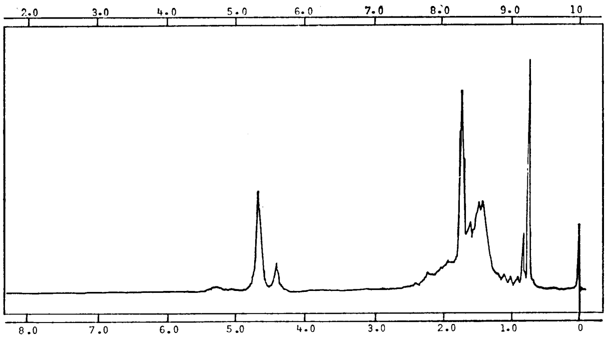

Fig. 1. NMR of β-selinene.

$$A = -\log T \tag{3}$$

$$A' = A/A_{\max} \times 100 \tag{4}$$

Here, T, A, and $A_{\max}$ stand for the apparent transmittance based on the base line of the spectral chart, the pseudo absorbance, and the intensity of the strongest band in the whole spectrum respectively. The numerals 1, 2, and 3 are given for the intensity (I) when $A' \leq 30$, $30 < A' \leq 70$, and $A' > 70$ respectively.

A line of numerals is given for the IR of β-selinene (101)(Fig. 2) as follows: 0 307 102 0 0 0 103 206 0 206 203 104 106 106 107 303 103 110. In numeral 307, the first figure 3 stands for the intensity and the last figure 7 the location of the strongest band in the division of 3200—2800. A band located just on the boundary is regarded as being in the higher division.

The mass spectrum was normalized by expressing it with the position and intensity of the two most intense peaks at 14 m.u. intervals, starting at m/e 40. To express the intensity, 1, 2, and 3 are given for the peak in accordance with its relative intensity, $\leq 35\%$, $>35\%$ and $\leq 75\%$, and $>75\%$, respectively. The mass spectrum of β-selinene (101)(Fig. 3) was converted into a numerically normalized state as follows: 413 131 532 552 672 691 793 813 912 933 1053 1073 1192 1212 1332 1351 1472 1491 1612 1751

** Present Address: Ube Industries Ltd., Ube, Yamaguchi 755.

*** Present Address: Toyohashi University of Technology, Tempaku, Toyohashi, Aichi 440.

TABLE 1. RESULTS GIVEN BY THE USE OF EITHER THREE TYPES OF SPECTRA, TWO, OR ONE ONLY (N, I, and M stand for the ^{1}H NMR, IR, and mass spectrometric data, respectively. The numeral in each column indicates the number of noise(s).[a])

No.	Compound	N	I	M	N+I	N+M	I+M	N+I+M
1	Acoradiene	0	5	1	0	0	0	0
2	Anethole	0	0	0	0	0	0	0
3	Angelic acid	0	0	1	0	0	0	0
4	Aromadendrene	2	0	7	0	0	0	0
5	Ascaridole	0	0	0	0	0	0	0
6	Auraptene	0	0	0	0	0	0	0
7	*β*-Bisabolene	0	4	2	0	0	0	0
8	Borneol	1	0	1	0	0	0	0
9	Bornyl acetate	2	0	1	0	1	0	0
10	Isobornyl acetate	3	0	1	0	1	0	0
11	Calamenene	0	0	0	0	0	0	0
12	Camphene	0	0	6	0	0	0	0
13	Camphor	2	2	2	1	0	0	0
14	3-Carene	0	2	4	0	0	1	0
15	*cis*-Carveol	1	4	0	1	0	0	0
16	*trans*-Carveol	0	4	0	0	0	0	0
17	*d*-Carvone	0	1	0	0	0	0	0
18	Carvotanacetone	2	0	1	0	0	0	0
19	Caryophyllene	0	0	8	0	0	0	0
20	*α*-Cedrene	2	3	1	0	0	0	0
21	Cedrol	0	1	0	0	0	0	0
22	*β*-Chamigrene	0	3	2	0	0	1	0
23	1,4-Cineol	0	3	1	0	0	0	0
24	1,8-Cineol	0	0	1	0	0	0	0
25	Citronellal	0	0	0	0	0	0	0
26	*β*-Citronellol	0	3	2	0	0	0	0
27	*α*-Copaene	2	2	0	0	0	0	0
28	Cuparene	0	0	0	0	0	0	0
29	*p*-Cymene	0	1	1	0	0	1	0
30	Dihydroactinidiolide	1	1	0	0	0	0	0
31	Dihydrocarvone	0	0	0	0	0	0	0
32	Diosphenol	0	0	0	0	0	0	0
33	*β*-Elemene	0	2	5	0	0	1	0
34	Elemol	0	0	4	0	0	0	0
35	Eucarvone	0	0	2	0	0	0	0
36	*β*-Eudesmol	2	0	0	0	0	0	0
37	Eugenol	0	0	1	0	0	0	0
38	Isoeugenol	0	0	1	0	0	0	0
39	Farnesol	0	1	0	0	0	0	0
40	Fenchone	0	0	0	0	0	0	0
41	Fenchyl alcohol	2	5	0	0	0	0	0
42	Geraniol	0	2	0	0	0	0	0
43	Guaiazulene	0	0	0	0	0	0	0
44	3-Hexen-1-ol	—	0	0	—	—	0	—
45	*α*-Humulene	0	0	0	0	0	0	0
46	Hydroxycitronellol	2	0	0	0	0	0	0
47	Isoimperatorin	0	0	0	0	0	0	0
48	*α*-Ionone	0	4	0	0	0	0	0
49	*β*-Ionone	0	1	0	0	0	0	0
50	*α*-Irone	1	0	0	0	0	0	0
51	Hinokitiol	0	0	0	0	0	0	0
52	Lavandulol	1	3	1	1	0	0	0
53	*α*-Limonene	0	5	4	0	0	0	0
54	Linalool	0	5	4	0	0	0	0

TABLE 1. (Continued)

No.	Compound	N	I	M	N+I	N+M	I+M	N+I+M
55	*cis*-Linalool oxide	0	0	2	0	0	0	0
56	*trans*-Linalool oxide	0	0	2	0	0	0	0
57	Linalyl acetate	0	2	3	0	0	0	0
58	Longifolene	0	1	6	0	0	0	0
59	Maltol	0	0	0	0	0	0	0
60	Mayurone	0	0	0	0	0	0	0
61	1,8-Menthanediol hydrate	0	0	0	0	0	0	0
62	Menthofuran	0	0	1	0	0	0	0
63	Isomenthol	0	3	4	0	0	2	0
64	1-Menthol	3	3	3	0	0	2	0
65	Neomenthol	1	2	3	0	0	2	0
66	Neoisomenthol	1	4	4	0	0	0	0
67	Menthone	2	4	1	1	0	0	0
68	Isomenthone	0	4	1	0	0	0	0
69	Methyl anthranilate	0	0	0	0	0	0	0
70	2-Hydroxy-3-methyl-2-cyclopenten-1-one	0	0	0	0	0	0	0
71	6-Methyl-5-hepten-2-one	0	3	2	0	0	0	0
72	Myrcenol	1	1	2	0	0	0	0
73	Myristicin	0	0	0	0	0	0	0
74	Myrtenal	0	0	2	0	0	0	0
75	Myrtenol	0	1	1	0	0	0	0
76	Nerol	0	3	10	0	0	0	0
77	Nerolidol	0	1	1	0	0	0	0
78	Nezukone	0	0	0	0	0	0	0
79	1-Nonene	0	0	0	0	0	0	0
80	Nootkatone	1	1	0	0	0	0	0
81	Occidentalol	0	0	0	0	0	0	0
82	Ocimene	0	0	5	0	0	0	0
83	1-Octen-3-ol	0	1	0	0	0	0	0
84	α-Patchoulene	2	1	0	0	0	0	0
85	β-Patchoulene	1	0	1	0	0	0	0
86	γ-Patchoulene	2	4	4	0	0	1	0
87	Patchouli alcohol	0	2	5	0	0	0	0
88	Perillaldehyde	0	0	3	0	0	0	0
89	Perillyl alcohol	0	1	3	0	0	0	0
90	α-Pinene	2	1	6	0	0	1	0
91	β-Pinene	0	2	9	0	0	0	0
92	Isopinocamphone	0	3	1	0	0	0	0
93	Piperitenone oxide	0	1	0	0	0	0	0
94	Piperitone	0	1	1	0	0	0	0
95	Piperonal	0	0	0	0	0	0	0
96	Pulegone	0	2	0	0	0	0	0
97	*cis*-Rose oxide	0	1	1	0	0	0	0
98	*trans*-Rose oxide	0	1	1	0	0	0	0
99	Sabinene hydrate	1	3	5	0	0	0	0
100	Safrole	0	0	0	0	0	0	0
101	β-Selinene	1	9	8	0	0	3	0
102	α-Terpinene	1	2	3	0	0	0	0
103	γ-Terpinene	1	1	1	0	0	1	0
104	Terpinen-4-ol	0	5	2	0	0	0	0
105	α-Terpineol	1	6	1	0	0	0	0
106	β-Terpineol	0	4	2	0	0	0	0
107	Thyjopsene	1	0	3	0	0	0	0
108	Thymol	0	0	0	0	0	0	0
109	Tiglic acid	0	0	1	0	0	0	0

Table 1. (Continued)

No.	Compound	N	I	M	N+I	N+M	I+M	N+I+M
110	Valeranone	0	0	0	0	0	0	0
111	Vanillin	0	0	0	0	0	0	0
112	Verbenone	0	4	0	0	0	0	0
	Maximum of noise	3	9	10	1	1	3	0
	Minimum of noise	0	0	0	0	0	0	0
	Sum total of noise	48	151	177	4	2	14	0
	Average	0.432	1.348	1.580	0.036	0.018	0.125	0.000

a) Noise stands for an inappropriate response for the sample taken.

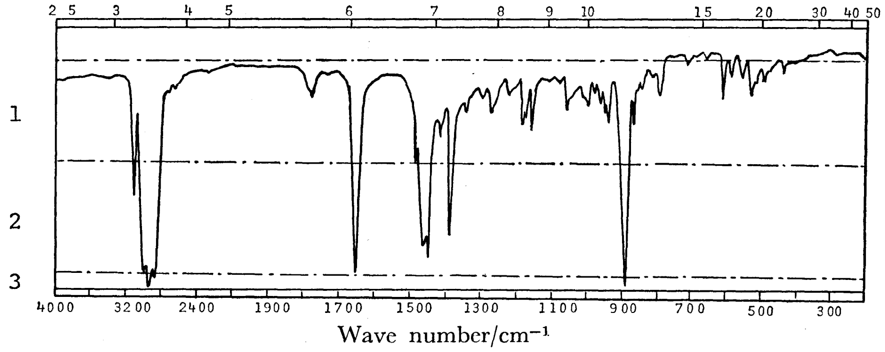

Fig. 2. IR of *β*-selinene.

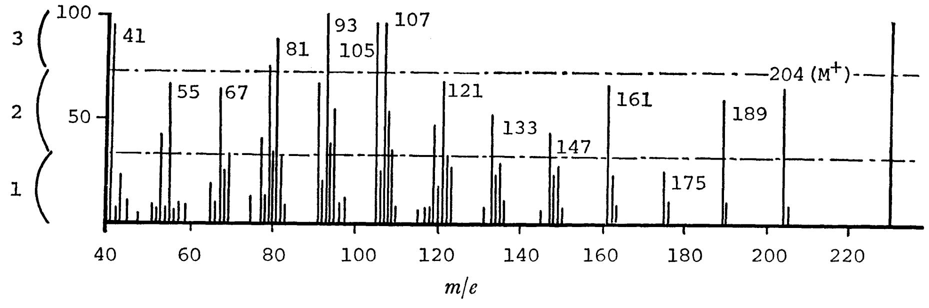

Fig. 3. MS of *β*-selinene.

1761 1892 1901 2042 2051. The last figure of each numeral indicates the intensity.

These normalized data of the NMR, IR, and MS of each compound are stored, with its code number (instead of compound name) in the computer.

Comparison of Spectra. On NMR comparison, when the conditions of $|G_y - G_x| \leqq 0.03$ and $|S_y - S_x| \leqq 0.04$ are satisfied, the x spectrum is judged to be equal to the y spectrum, where G_x and G_y mean the G values, and where S_x and S_y mean the S values, of the x and y spectra, respectively.

A parameter (E) defined by Eq. 5 is used for the IR comparison, where P_{xi} and P_{yi} mean the position of the strongest band in the ith main division of a

$$E = \sum_i |P_{xi} - P_{yi}| \qquad (5)$$

spectrum registered in the file, and that of the spectrum to be compared, respectively. When two spectra were compared, another parameter (M) was also used; this parameter represents the sum of the main divisions which have similar P values but different I values. As the most appropriate value, 20 for E and 6 for M are chosen by trial and error.

Two kinds of parameters (D and Q) are used in the MS comparison. D is the number of peaks which have a similar mass but a different intensity in the comparison of an unknown with each of the stored compounds. Q is the number of peaks which are lacking in the spectrum of an unknown but which are present in that of to be compared. The spectra of 10 compounds collected arbitrarily were measured at several different conditions in order to decide the most appropriate threshold values of D and Q. As a result, it was observed that D is distributed within the range from 0 to 3, and Q, within that from 0 to 4. Thus, taking a sufficient allowance into account, 5 and 7 were selected as the threshold values of D and Q, respectively. When D and Q are observed to be below the thresholds in the sorting upon MS, the two spectra are regarded as identical.

Results

The data base consists of terpenes and related compounds. They are in a structurally close relationship with each other, and a similar type of pattern is observed for their spectra in many cases. If sorting can discriminate one spectrum from all the other spectra in the system containing such kinds of compounds, the method may be considered to be a good tool for structure identification. Indeed, the sorting by the combined spectrometric data of NMR, IR, and MS gave excellent results.

Table 1 shows the sorting results afforded by the use of either all three types of spectra, of two, or of one only. As is shown in the table, one completely correct compound is retrieved by sorting either NMR, IR, or MS alone for almost one-third (2, 5, 6, 11, 25, 28, 31, 32, 40, 43, 44, 45, 47, 51, 59, 60, 61, 69, 70, 73, 78, 79, 81, 95, 100, 108, 110, and 111) of the registered compounds. The remaining two-thirds are contaminated with noise(s) when sorted with a single set of spectrometric data. However, the number of these compounds accompanied by contamination rapidly diminishes by sorting with either two or three types of spectra. For example, compound 101 is contaminated with one (145), nine (3, 15, 31, 33, 45, 67, 107, 173, and 183), and eight (9, 39, 45, 67, 117, 173, and 175) noises by sorting with only NMR, IR, or MS respectively. However, sorting with two types of spectra (NMR+IR, NMR+MS) gives no noise, though sorting with IR and MS gives three noises (45, 67, 173). Furthermore, as is shown in the most right column of Table 1, when three types of spectra are used, only one correct answer is always obtained for all compounds. The fact that nothing was retreived when three types of spectra of α-pseudowiddrene, which had not yet been registered in the data base, were compared with the store also proves the authors' method to be completely reliable.

References

1) R. A. Sparks, "Storage and Retrieval of WYANDOTTE-ASTM Infrared Spectral Data Using an IBM 1401 Computer," American Society for Testing Materials, Philadelphia, Pa. (1964).

2) D. H. Anderson, and G. L. Covert, *Anal. Chem.*, **39**, 1288 (1967).

3) D. S. Erley, *Anal. Chem.*, **40**, 894 (1968).

4) F. E. Lytle and T. L. Brazie, *Anal. Chem.*, **42**, 1532 (1970).

5) G. A. Massios, *Am. Lab.*, **3**, 55 (1971).

6) R. W. Sebasta and G. G. Johnson, Jr., *Anal. Chem.*, **44**, 260 (1972).

7) C. S. Rann, *Anal. Chem.*, **44**, 1669 (1972).

8) J. Zupan, D. Hadzi, and M. Penca, *Kem. Ind.*, **23**, 275 (1974).

9) E. C. Penski, D. A. Padowski, and J. B. Bouck, *Anal. Chem.*, **46**, 955 (1974).

10) K. Schaarschmidt, R. Riemer, and E. Steger, *Z. Chem.*, **14**, 374 (1974).

11) H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, *J. Chem. Inf. Comput. Sci.*, **15**, 207 (1975).

12) R. C. Fox, *Anal. Chem.*, **48**, 717 (1976).

13) J. Zupan, D. Hadzi, and M. Penca, *Comput. Chem.*, **1**, 71 (1976).

14) E. M. Kirby, R. N. Jones, and D. G. Cameron, *CODATA Bull.*, **21**, 18 (1976).

15) J. Zupan, J. T. Clerc, and D. Hadzi, *Vestn. Slov. Kem. Drus.*, **23**, 73 (1976).

16) B. Pettersson and R. Ryhage, *Ark. Kemi.*, **26**, 293 (1967).

17) L. R. Crawford and J. D. Morisson, *Anal. Chem.*, **40**, 1464 (1968).

18) S. L. Grotch, *Anal. Chem.*, **42**, 1214 (1970).

19) B. A. Knock, I. C. Smith, D. E. Wright, R. G. Ridley, and W. Kelly, *Anal. Chem.*, **42**, 1526 (1970).

20) H. S. Hertz, R. A. Hites, and K. Beiemann, *Anal. Chem.*, **43**, 681 (1971).

21) S. L. Grotch, *Anal. Chem.*, **43**, 1362 (1971).

22) L. E. Wangen, W. S. Woodward, and T. L. Isenhour, *Anal. Chem.*, **43**, 1605 (1971).

23) S. R. Heller, "DCRT/CIS, MSSS User's Manual," Division of Computer Research and Technology, Bethesda, Md. (1972).

24) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972).

25) S. R. Heller, H. M. Fales, and G. W. A. Milne, *Org. Mass Spectrom.*, **7**, 107 (1973).

26) S. L. Grotch, *Anal. Chem.*, **45**, 2 (1973).

27) P. R. Naegeli and J. T. Clerc, *Anal. Chem.*, **45**, 739A (1974).

28) R. Schwarzenbach, J. Meili, H. Konitzer, and J. T. Clerc, *Org. Magn. Res.*, **8**, 11 (1976).

29) R. J. Feldman, S. R. Heller, K. P. Shapiro, and R. S. Heller, *J. Chem. Doc.*, **12**, 41 (1972).

30) J. T. Clerc, C. Jost, T. Meier, and R. Schwarzenbach, *Chimia*, **27**, (12) 665 (1973).

31) J. T. Clerc and F. Erni, *Top. Curr. Chem.*, **39**, 91 (1973).

32) J. T. Clerc, "Computers in Chemical Research and Education, Proceedings," D. Hadzi, ed, Elsevier Publishing Co., Amsterdam (1973), Vol. 2, pp. 3/109.

33) L. A. Gribov, V. A. Dementyev, M. E. Elyashberg, and E. Z. Yakupov, *J. Mol. Struct.*, **22**, 161 (1974).

34) V. A. Koptyug, *Z. Chem.*, **15**, 41 (1975).

35) N. A. Gray, *Anal. Chem.*, **47**, 2426 (1975).

36) The NIH-EPA Chemical Information Systems, Status Report No. 4, Dec., 1976.

37) J. Zupan, M. Penca, D. Hadzi, and J. Marsel, *Anal. Chem.*, **49**, 2145 (1977).

38) S. Sasaki, Y. Kudo, H. Abe, and T. Yamasaki: "Computer-assisted Structure Elucidation," ACS Symposium Series, (1977), No. 54, p. 108.

39) "Spectral Atlas of Terpenes and the Related Compounds," ed by Y. Yukawa and Sho Ito, Hirokawa Publishing Co., Tokyo (1973).

40) S. Sasaki, Y. Yotsui, and S. Ochiai, *Bunseki Kagaku*, **24**, 213 (1975).