SPECTRA: A Spectral Information Management System Featuring a Novel Combined Search Function

Hideyuki Masui*

Sumitomo Chemical Company, Organic Synthesis Research Laboratory,
Tsukahara, Takatsuki-shi 569-11, Japan

Mototsugu Yoshida

Sumitomo Chemical Company, Tsukuba Research Laboratory,
Kitahara, Tsukuba-shi 300-32, Japan

Received September 6, 1995⊗

The SPECTRA collection of software as a spectral information management system for organic compound structure determination is described. The SPECTRA (SPECTral Research and Analysis) system suggests candidate structures for chemical compounds based on analysis of their spectra, where mass spectra, infrared spectra, ^{1}H-nuclear magnetic resonance spectra, and ^{13}C-nuclear magnetic resonance spectra are possible input. The system computes the optimal matching of an input spectrum with stored spectra in a database and also retrieves the spectra of compounds that contain a substructure of the unknown compound. A novel combined search algorithm can be activated when two to four spectra are given as information of an unknown compound. Similarities between the input spectrum and each spectrum in the database are calculated, and the corresponding candidate compounds are ranked according to their similarity score.

INTRODUCTION

Chemists in such diverse fields as electronics, biotechnology, and new materials are constantly synthesizing new compounds, but only a few out of tens of thousands will actually reach the market and contribute to social advancement with novel superior properties. However, development of improved and especially new products is vital for the survival of chemical industries, and structure determination of conceived candidates is an important part of the development process. Computational chemistry has become an indispensable tool for solving these real world problems. The computational chemistry approach comprises three categories: molecular modeling, organic synthesis design, and structure elucidation. We describe here a spectral information management system for organic compounds named SPECTRA that has been developed for improving the efficiency of the research and development process by aiding in the structure elucidation step. It employs a precise retrieval method on a database containing many data related to one’s field of interest. The system stores and retrieves four kinds of spectral information and the corresponding chemical structures. A novel combined search function included in this system consistently provides more reliable results by using several kinds of spectral data simultaneously.

SYSTEM DESCRIPTION

1. Overview. SPECTRA[1,2] is a computational system to be used by chemists when determining the structures of organic compounds. It has been jointly developed by Sumitomo Chemical Company and NEC Corporation for the following main purposes: (1) construction of in-house databases of spectral data and the corresponding chemical structures, (2) combination of data from external vendors with in-house data, and (3) enhancement of efficiency in structure determination. The SPECTRA system has the ability to combine information from mass spectra, infrared spectra, proton magnetic resonance spectra, and carbon 13 magnetic resonance spectra to determine the closest matching compounds. This gives a significant advantage over systems which are designed for only a single type of spectra[3−8] as will be demonstrated in a later section.

Figure 1 shows the system configuration of SPECTRA. It has a storage mode and a retrieval mode. The storage mode is designed for registration of chemical structures and spectral data. In this mode the system requires input spectral data for which the corresponding structures have already been solved. The retrieval mode is for the search of chemical structures and up to four kinds of spectra. It performs optimal matching of an input spectrum with those in the database. Chemical structures containing a substructure of the input can also be retrieved. In addition, the system can graphically display the chemical structures and spectra retrieved by the search process.

2. Data Base. A database consists of a chemical structure file and the corresponding spectral pattern file from which the spectral data with the four different kinds of spectra are derived. The chemical structure file includes ID numbers, chemical compound names, molecular formulas, connection tables, and so forth. The spectral data file contains codes, peak data, and several properties of the spectra which are key items in the retrieval process. Examples of these keys are lists of peaks with normalized intensities (0−1000), “rectangular arrays” [4] for the mass spectra, the codes for ^{13}C-NMR spectra, multiplicities for the NMR peaks, the chemical shift of peaks with the maximum intensity in the NMR spectra, and so forth.

The elements of the rectangular arrays are, in descending order, 14 sums of the intensities of the ions at m/e $6+14n$, $7+14n$, ... $19+14n$, where $n = 0, 1, 2...$ The mass sequences

⊗ Abstract published in *Advance ACS Abstracts,* February 1, 1996.

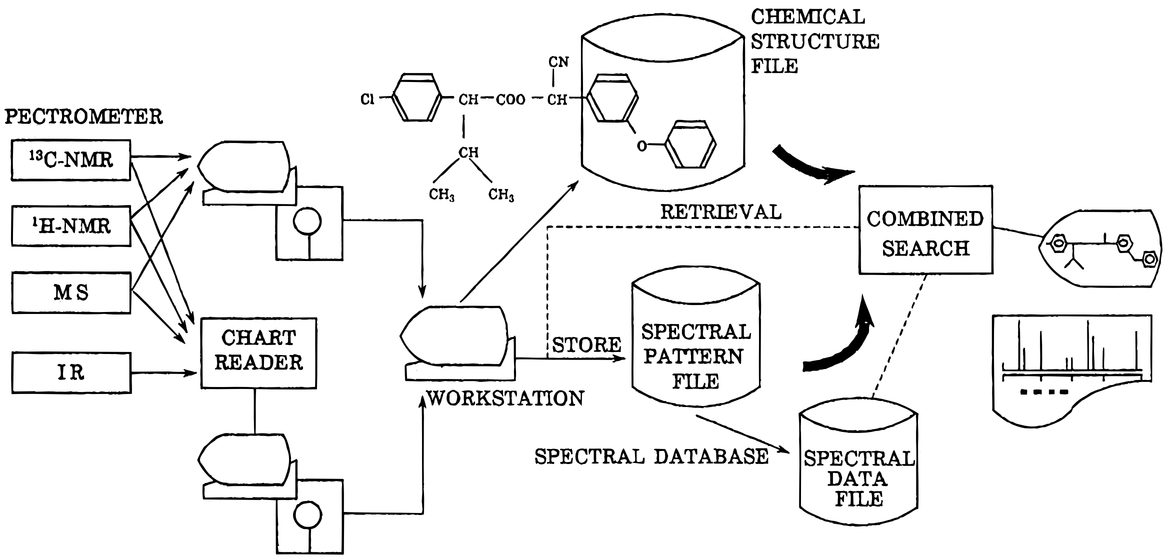


Figure 1. The SPECTRA system configuration.

represented by the six highest sums are converted into binary codes for speed of comparison. In many cases, several relative intensities of the ions are obtained depending on measurement conditions. Several different permutations within certain constraints for the rectangular arrays are produced and registered. It makes it possible to compensate for the different conditions and carry out an effective presearch.

The full range of a ^{13}C -NMR spectrum (−100 ppm to 300 ppm relative to TMS) is divided into 16 nonequidistant intervals.[9] Example of these are $50\ \text{ppm} < X_6 \leq 65\ \text{ppm}$, $65\ \text{ppm} < X_7 \leq 82\ \text{ppm}$, $82\ \text{ppm} < X_8 \leq 105\ \text{ppm}$, where X_i is the interval range, $i = 1, 2, ..., 16$. The number of peaks in each interval is encoded and each multiplicity (quartet, triplet, doublet, singlet, unknown) is encoded for by 1 (presence) or 0 (absence) for each interval. These codes are used for the presearch of ^{13}C-NMR spectra.

The spectral pattern file includes data concerning the range of each spectrum and the corresponding signal intensities to be used when creating the spectral image display. Digital resolutions of the spectra are 0.001 ppm for the ^{1}H-NMR and 0.01 ppm for the ^{13}C-NMR. The IR spectra have two different resolutions according to the wave number. The digital resolutions are 2 cm^{-1} for the range between 4000 and 2000 cm^{-1} and 1 cm^{-1} between 2000 and 400 cm^{-1}. These data are compressed and stored in the database.

At present, the system comprises data for about 132 000 compounds with about 150 000 spectra in the databases. The spectral data consists of 71 000 mass spectra, 27 000 IR spectra, 46 000 ^{13}C-NMR spectra, and 7000 proton NMR spectra. Several different databases can be constructed with this system for specific purposes. For example, one could be for a company-wide usage, and another could be reserved for the section being the owner of the database.

3. Storage Mode. Substructures or full molecular structures are easily drawn with some simple manipulations on the graphic display using a mouse device. Templates of symbol strings have been prepared such that the desired structure can easily be drawn. Upon completion, the structure can be registered in a chemical structure file. The system also accepts structures written in some CT (Connection Table) formats.

Concerning the spectral data, the system accepts spectral patterns from spectrometers, strip charts, spectral data supplied by external vendors, and so forth. The spectral data are stored in the spectral pattern file and the corresponding spectral data file. The system requires the input of the multiplicity of each peak in the NMR spectra. The multiplicities of a ^{1}H-NMR spectrum come from its assigned data and those of a ^{13}C-NMR spectrum come from information of INEPT (Insensitive Nuclei Enhanced by Polarization Transfer) or DEPT (Distortionless Enhancement by Polarization Transfer) spectra. The spectral information and the corresponding structures are connected with ID numbers. However, the system not only handles four kinds of spectra but also additional spectral taken under different measurement conditions. For instance, the system allows registration of two mass spectra of the same compound that are measured by, for example, the ionization conditions of EI (Electron Impact) method and FD (Field Desorption) method, respectively.

4. Retrieval Mode. The system is able to retrieve several key items in the retrieval mode such as chemical compound names, molecular formulas, molecular weights, and so forth. In addition, the system retrieves compounds containing the unknown structure as part of their structure. It is also possible to retrieve database spectra and molecular structure information from input spectra. In this mode, there are two ways to input spectra. One is to type in the peak data manually. The other way is to input a spectral pattern into the system using utilities described later. In order to reduce the solution space to a manageable number of candidates the system retrieves codes for an ionization method for mass spectroscopy, a magnetic field code for ^{1}H-NMR spectroscopy, and so forth. The SPECTRA system extracts features of the peak data from each input spectrum. These features are spectral parameters, peak codes, peak list, no signal band, and so forth. Depending on the input data the system carries out the appropriate processing. For instance, when a mass spectral pattern of an unknown compound is supplied as input, the system calculates the rectangular array and uses it for the presearch. For the ^{13}C-NMR spectrum of the unknown compound the codes for the number of peaks and the multiplicities are used in the presearch.

Table 1. Parameters for the Calculation of Similarity Score

spectrum	parameters
MS	*m*/*z* values, intensities
IR	wave numbers, intensities
^{1}H-NMR	chemical shifts, multiplicities, intensities
^{13}C-NMR	chemical shifts, multiplicities

The system retrieves spectra candidates from the database after completion of the presearch. All peaks in the spectra are matched to those in the database within some margins except for the mass spectral peaks. When multiplicities for the NMR spectrum are supplied, the system matches a peak from the database with the same multiplicity to the input peak. If multiplicities of the input spectrum are unknown, the system can handle the data without them as well.

The system calculates the similarity score between each spectrum in the database and the input spectrum upon completion of the spectra retrieval processing. Parameters in Table 1 will be used for the calculation of these scores. The system has several equations for calculating the similarity scores for the four kinds of spectra. Two of the equations resemble the calculation of correlation coefficients by using associated intensities with matched *m*/*z* values for mass spectra or wave numbers in case of infrared spectra. For the NMR spectra multiplicities are given priority in matching the peaks and are also used in calculation of the similarities. The intensities in the ^{1}H-NMR spectra which are converted into number of protons are also used in the score calculation. The scores are calculated not only for matched peaks but also for unmatched peaks as a negative score. Empirically matched factors (0.3−1.0) are used as a basis for the calculation of scores for the multiplicity and for the proton number.

If two to four spectra of an unknown compound are given as input for the retrieval process, the system can carry out the combined search. The details will be described in the next section. The overall scores of the resulting candidate compounds are the averages of the similarities for each kind of spectrum. The results are ranked in order of the total scores.

5. Combined Search. An early paper[10] has stated that combination of several different spectroscopic methods can greatly enhance the performance. Although a few data of useful reference compounds were missing in databases, an integrated system for comparing spectral data should be able to handle incomplete data. However, they did not give any concrete information on the nature of the incomplete data. In many systems employing the combined search[11−16] the missing data problem is not mentioned, or it is premised that each compound has all of the corresponding spectra necessary for a test of the combined search. As J. T. Clerc et al.[10] pointed out, it is usually the case that a spectrum or a few spectra for a number of compounds are missing in databases. In this case their algorithm may miss the best solution. We show later that the novel search method in SPECTRA is able to obtain the best result in a realistic case.

For instance, a compound (tentatively called compound-A) has the MS, ^{1}H-NMR, and ^{13}C-NMR spectra in the database, but the IR spectrum is missing. When the four kinds of spectra of the unknown compound taken, to be identical to the compound-A, are input as query spectra, the compound-A is discarded from a list of candidates in the search sequence of the IR spectrum. Then because this list usually serves as input for the subsequent retrieval, the system cannot obtain the compound-A in the final list. Whatever matching compounds are obtained from the search sequence based on the other spectra, it will be rejected from the final list since the IR spectrum search is included automatically in the search sequence depending on the query spectra. The same result is obtained by the logical operation method for the combined search.

The novel combined search method described in this paper does not miss the correct answer in the missing spectrum case explained above. The first step in the novel combined search is that the system retrieves spectral information in the database using the input peak information. Three categorical sets result from this retrieval process. The first is a list of compound spectra being similar to the query one (A set). The second is a list of compounds for which the same kind of spectrum as the input has not been registered (B set). The third is a list of compound spectra which do not match the input (C set). The C set is not used in the subsequent process. The A and B sets are compiled to serve as a source list for the second step. Similar procedures are repeated for the other remaining spectra. Even if the compounds have only one spectrum in the database, they should remain in the solution space. It will eventually provide the chemists with good hints for the structure elucidation. In this way the combined search method effectively uses the available data in the database for practical structure determination.

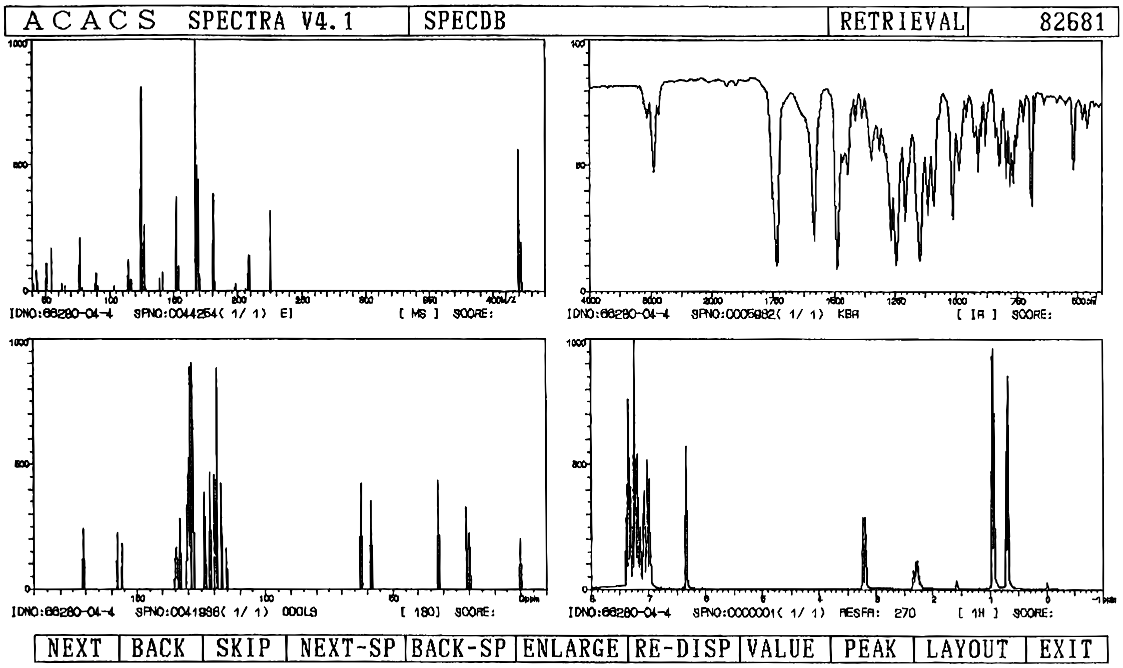
The next step in the structure determination process is calculation of similarity scores for the retrieved spectra and subsequent ranking of candidate compounds. If the overall similarity scores for two candidate compounds coincide the one which has the larger number of accorded spectra will be ranked higher in the list.

6. Example. To illustrate the effectiveness of this combined search, an example is shown below. Suppose the four spectra of methyl octadecanoate (CA Registry number: 112-61-8) are used as input for the spectral data of an unknown compound. Using all spectra which are supplied as spectral pattern data the combined search provides the first nine top rated compounds, where the number depends on a threshold level for similarity scores (Table 2). The correct structure ranks first in the list, and hence it is easy to recognize the most appropriate one. The list includes some compounds for which only one spectrum has been registered in the database. In Table 2 the compounds ranked no. 4 and no. 8 have only mass spectra recorded, but both are similar in structure to the correct compound showing the advantage of the combined search. According to the result, the unknown compound includes a long chained alkyl group and an ester group. All compounds in the candidate list are seen to be methyl esters. Even if the database does not have the target compound, which would appear as no. 1, it is a simple matter to estimate the correct structure from the list.

7. Display. The SPECTRA system graphically displays both structures and spectra and offers the possibility to have from one and up to four pictures on a screen. Figure 2 shows a screen image of the graphic display of four measured spectra. The system is also able to display two spectra together, superimposed for visual comparison. It is furthermore possible to “zoom” into a part of a spectrum by using

Table 2. The Result of the Combined Search

no.	structure	similarity score MS	IR	CMR	HMR	total
1	$CH_3-(CH_2)_{16}-CO-O-CH_3$	100	100	100	100	100
2	$CH_3-(CH_2)_{17}-CO-O-CH_3$	69	98	96	100	91
3	$CH_3-(CH_2)_{15}-CO-O-CH_3$	66	91	96	100	88
4	$(CH_3)_2{>}CH-(CH_2)_{14}-CO-O-CH_3$	87				87
5	$CH_3-(CH_2)_{18}-CO-O-CH_3$	70	96	91	100	86
6	$CH_3-(CH_2)_{22}-CO-O-CH_3$	62	91	86	97	84
7	$CH_3-(CH_2)_{20}-CO-O-CH_3$	68	92	85		82
8	$CH_3-(CH_2)_{19}-CO-O-CH_3$	73				73
9	$CH_3-(CH_2-CH{=}CH)_3-(CH_2)_7-COO-CH_3$		83		61	72



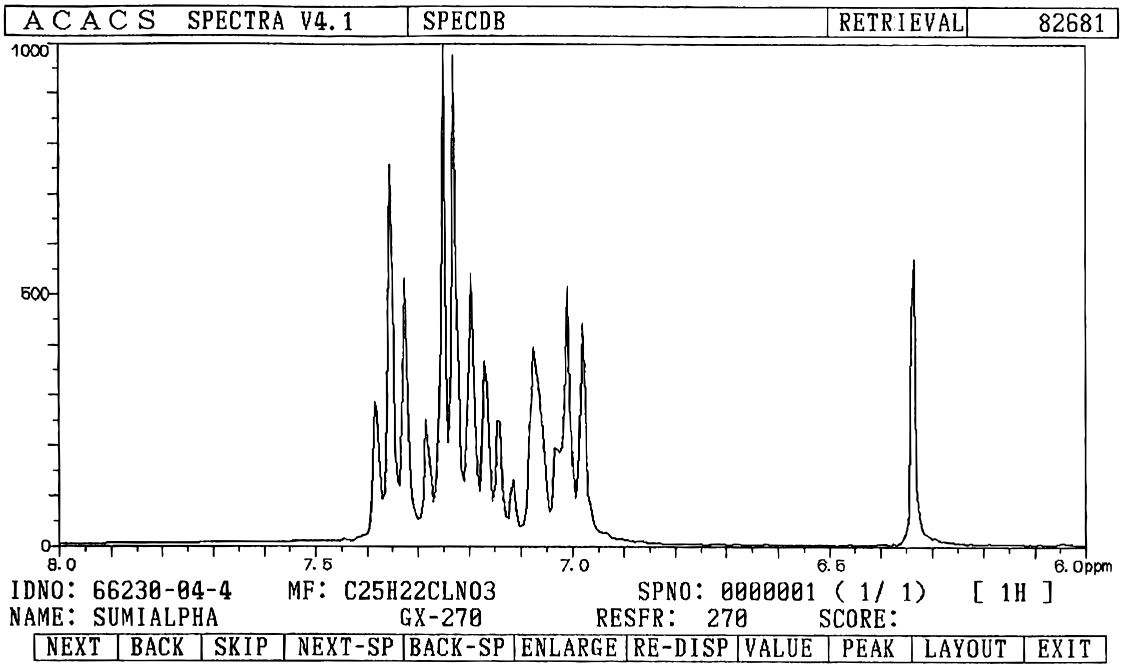
IDNO:66280-04-4 SPNO:0044254(1/ 1) EI [MS] SCORE:

IDNO:66280-04-4 SPNO:0005982(1/ 1) KBR [IR] SCORE:

IDNO:66280-04-4 SPNO:0041998(1/ 1) CDCL3 [13C] SCORE:

IDNO:66280-04-4 SPNO:0000001(1/ 1) RESFR: 270 [1H] SCORE:

Figure 2. A screen image of the graphic display.



IDNO: 66230-04-4 MF: C25H22CLNO3 SPNO: 0000001 (1/ 1) [1H]
NAME: SUMIALPHA GX-270 RESFR: 270 SCORE:

Figure 3. A screen image of an enlarged ^{1}H-NMR spectrum.

the "ENLARGE" subcommand. Figure 3 shows an example of an enlarged ^{1}H-NMR spectrum. The Y axis of the screen for the spectrum gives the normalized intensity of the signal. The X axis represents the *m*/*z* for mass spectra, the wave number for infrared spectra, and the chemical shift for NMR spectra, respectively.

8. Utility. Some utility programs have been developed for supporting the SPECTRA system. A WLN translation utility is used for conversion from WLN (Wiswesser Line-Formula Chemical Notation) to the corresponding connection table. Database conversion utilities are employed for transcribing spectral files from commercially available databases, such as Wiley/NBS for mass spectra, INKA for ^{13}C-NMR spectra, SDBS for infrared, proton NMR, and ^{13}C-NMR spectra, and so forth. Spectral pattern input utilities have been designed for handling the input of spectral patterns directly from several kinds of spectrometers to the SPECTRA system. The spectral patterns can also be read in using an optical chart-reading instrument named “CHART READER”.[17] The CHART READER consists of two sections. One is a reader section, and the other is a personal computer section for data processing. The reader section comprises a CCD camera, a camera controller, a run-length data circuit, a movable table on which a strip chart recorded spectrum is placed, a stepping motor, and a motor controller. The resolution of the stepping motor is 20 nm. Maximum speed is 100 mm/s. In the reader section, the CCD camera photographs a spectrum image from a strip chart as run-length data, and the computer converts the image data into digital data. The system is able to deal with all four kinds of spectra. Peak picking procedures are automatically achieved for the mass, IR, and ^{13}C-NMR spectra, but the system needs help by a chemist in order to recognize group of peaks or multiplicities in the ^{1}H -NMR spectra. The final data are put onto a diskette that is used for the registration of the spectral patterns.

IMPLEMENTATION

The SPECTRA suite of programs and databases have been installed in an NEC system 3400 computer running under the ACOS operating system. The programs are written in COBOL, FORTRAN, and CPL (Common Programming Language developed by NEC). For terminals we use NEC N5200 graphic workstations and NEC personal computers PC-9821. The SPECTRA system can be used simultaneously and in an interactive mode from each connected terminal. A high speed digital communication network has been established for all our factories and research laboratories throughout the country, and local area networks in some research laboratories are set up. The SPECTRA system is used through this network.

CONCLUSION

The SPECTRA system performing storage and retrieval of chemical structures and molecular spectra has developed into a state-of-the-art software for organic compound structure determination. The main features of the system are as follows:

(1) Novel combined search for different types of spectra.

(2) Optimal matching of an input spectrum with spectra in the data base.

(3) Retrieval of the spectra of compounds which contain a substructure of the input.

(4) Handling of four kinds of spectra (mass spectra, infrared spectra, proton NMR spectra, and carbon 13 NMR spectra) and chemical structures.

(5) Display of the closest matching spectra together with those of the unknown substances.

(6) Combination of data from external vendors with in-house data.

We can conclude that the SPECTRA system enhances efficiency and productivity in the determination of chemical structures for improved research and development in the field of organic chemistry.

REFERENCES AND NOTES

(1) Yoshida, M.; Masui, H. Spectral Information Management System—SPECTRA. *Sumitomo Kagaku* **1988**, (1), 54−62.

(2) Masui, H.; Yoshida, M.; Moriguchi K.; Shigenaga, H.; Yokota, A.; Mizobe, Y. A Computer System for Structure Determination of Organic Compounds—SPECTRA. *Proceedings of the 25th Annual Meeting on Information Science and Technology* **1988**, 237−245.

(3) Knock, B. A.; Smith, I. C.; Wright, D. E.; Ridley, R. G.; Kelly, W. Compound Identification by Computer Matching of Low Resolution Mass Spectra. *Anal. Chem*. **1970**, *42*, 1516−1520.

(4) Hertz, H. S.; Hites, R. A.; Biemann, K. Identification of Mass Spectra by Computer-Searching a File of Known Spectra. *Anal. Chem*. **1971**, *43*, 681−691.

(5) Heller, S. R. Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination. *Anal. Chem*. **1972**, *44*, 1951−1961.

(6) Grotch, S. L. Computer Identification of Mass Spectra Using Highly Compressed Spectral Codes. *Anal. Chem*. **1973**, *45*, 2−6.

(7) Naegeli, P. R.; Clerc, J. T. Computer System for Structural Identification of Organic Compounds from Spectroscopic Data. *Anal. Chem*. **1974**, *46*, 739A−744A.

(8) Schwarzenbach, R.; Meili, J.; Koenitzer, R.; Clerc, J. T. A Computer System for Structural Identification of Organic Compounds from ^{13}C NMR Data. *Org. Magn. Reson*. **1976**, *8*, 11−16.

(9) Novič, M.; Zupan, J. Hierarchical Clustering of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chim. Acta* **1985**, *177*, 23−33.

(10) Clerc, J. T.; Erni, F. Identification of Organic Compounds by Computer-Aided Interpretation of Spectra. *Top. Curr. Chem*. **1973**, *39*, 91−107.

(11) Gray, N. A. B. Structural Interpretation of Spectra. *Anal. Chem*. **1975**, *47*, 2426−2431.

(12) Zupan, J.; Penca, M.; Hadži, D.; Marsel, J. Combined Retrieval System for Infrared, Mass, and Carbon 13 Nuclear Magnetic Resonance Spectra. *Anal. Chem*. **1977**, *49*, 2141−2146.

(13) Sasaki, S.; Abe, H.; Saito, K.; Ishida, Y. The Computer-assisted Characterization of Terpenes and Related Compounds by the Use of Combined Spectrometric Data. *Bull. Chem. Soc. Jpn*. **1978**, *51*, 3218−3222.

(14) Tanabe, K.; Hiraishi, J.; Tamura, T.; Yamamoto, O.; Yanagisawa, M.; Wasada, N. Efficiency of the Combined Search by Infrared, 13C-NMR, and Mass Spectral Data. *Bunseki Kagaku* **1984**, *33*, 95−98.

(15) Tanabe, K.; Hiraishi, J.; Tamura, T.; Yamamoto, O.; Yanagisawa, M.; Wasada, N. COSMOS-Combined Search System for Molecular Spectra. *Comput. Enhanced Spectrosc*. **1984**, *2*, 97−99.

(16) Yamamoto, O.; Someno, K.; Wasada, N.; Hiraishi, J.; Hayamizu, K.; Tanabe, K.; Tamura, T.; Yanagisawa, M. An Integrated Spectral Data Base System Including IR, MS, ^{1}H-NMR, ^{13}C-NMR, ESR and Raman Spectra. *Anal. Sci*. **1988**, *4*, 233−239.

(17) Yoshida, M.; Moriguchi, K.; Shigenaga, H.; Masui, H.; Yokota, A.; Mizobe, Y. *Abstracts of Papers, 10th Symposium on Chemical Information and Computer Science, Nagoya* **1986**, 194−195.

CI950111V