AN EFFICIENT SEARCH PROGRAM FOR ASTM INFRARED SPECTRAL DATA

Kazutoshi TANABE, Tadao TAMURA, Jiro HIRAISHI, and Shinnosuke SAËKI
National Chemical Laboratory for Industry, Tsukuba, Ibaraki 305

A computer search system has been proposed whose characteristic consists in the high efficiency of search due to the use of peak intensity data in the matching process and in the reduction of search time due to the use of a subfile system. A test search using 100 measured spectra as a sample proved the excellent performance of the present search system.

1. Introduction

Numerous computer search programs have been written for the ASTM infrared data file,¹⁻¹³⁾ and some of those systems are being used on the commercial base. The ASTM data do not contain any information on peak intensity (except for masking if a group of peaks includes one or more strong peaks). Thus, data of peak intensities in a sample spectrum are not utilized effectively in the matching process of the so far proposed search systems. However, the utilization of peak intensity data of a sample spectrum in the matching process of the search program will improve the efficiency of search.

The ASTM file has widely been used because, in spite of the ambiguity in the coding of spectra, the file already includes spectral data for more than 100,000 compounds enough for practical purposes. In order to reduce the search time for such a large data base, chemical information (molecular formula, functional group, melting or boiling point) other than infrared spectral data may be entered as the input data for search. But, such chemical information is not always given for a test sample, and therefore the use of subfiles created from the ASTM mother file will be effective in reducing the search time.

From these viewpoints, we developed a new search program for the ASTM file, and tested its performance in this study. In this paper, we report the outline of our system and the result of its test search. The characteristic of the system consists in the high efficiency of search by utilizing peak intensity data of a sample spectrum in the matching process, and in the reduction of the search time by using subfiles created from the ASTM mother file.

2. Methods

2.1 Input Data

In the present search system, frequencies and intensity classes for peaks, and frequency ranges for no-peak areas, are read from a sample spectrum, and entered as the input data. The wavenumber of the maximum absorption is entered for the peak

frequency, and in the program, the appropriate allowance as described below is calculated for the matching process. As the input data of the no-peak areas, the wavenumbers of two peaks lying on either side of the area where no peak is found in the sample spectrum are entered, and in the program, the area is reduced by the allowance of the peak frequencies of both peaks. If the input data of the no-peak areas become meaningless after the procedure, the data are automatically excluded from the input list.

2.2 Intensity Class

All peaks found in the sample spectrum are classified into five groups according to their relative peak intensities. Intensity class is determined by roughly estimating the absorbance of a peak and normalizing to the strongest peak. The strongest peaks are given the class 5, and the weakest the class 1. An example of the input data is shown in Fig. 1 and the sample spectrum is in Fig. 2. In the matching process of the search program, the input data of the class 5 peaks and of the no-peak areas are treated as a mandatory condition (that is, only spectral data in the ASTM file satisfying the condition are considered as "hit" and their scores are calculated), and other peak data are treated as a desirable condition.

INPUT DATA

PEAK 28

3110-3 1963-2 1830-1 1694-5 1609-4 1528-5 1488-1 1432-3 1403-3 1383-2
1360-3 1344-5 1319-4 1294-2 1260-5 1243-3 1113-3 1104-2 1077-1 1027-1
1007-3 992-1 977-1 963-3 857-5 820-1 748-4 693-3

NO-PEAK 11

2850-2470 2350-1963 1963-1830 1830-1694 1609-1528 1243-1113 1077-1027
963- 857 820- 748 748- 693 693- 670

Fig. 1. An example of input data (sample: p-nitroacetophenone)

Fig. 2. Infrared spectrum of p-nitroacetophenone

2.3 Allowance of Peak Position

The input data of peak frequencies are converted on the wavelength scale because the coding of the ASTM data is based on the scale, and the allowance value is computed for the peak position for considering peak shifts between the sample and reference spectra. In this study, the allowance value has been determined on the experimental basis, and for that purpose, peak shifts between the sample and reference spectra for the test compounds have been compared. From those data, the allowance value of the peak position has been determined as

$$\Delta\lambda = 0.0023 \times (\lambda_{\max} - 7)^2 + 0.08$$

where $\lambda_{\max}$ is the peak wavelength in μm .

2.4 Score

The spectral data of each member in the ASTM file are checked by examining whether the mandatory condition is fulfilled or not, taking the peak allowance values into account, and for "hit" members satisfying all of the mandatory condition, the score measuring the degree of coincidence is computed. In the present search program, the score is computed from the equation

$$S = 100 \times \left[\sum_{i=p+1}^5 (i \cdot N_i) - \sum_{i=1}^{p-1} (p-i)^2 \cdot N_i \right] / \left[\sum_{i=p+1}^5 (i \cdot N_i) \right]$$

where N_i is the number of hit peaks of the i -th intensity class, and p the highest intensity class where the number of hit peaks is smaller than that of the input peaks. When a peak of some intensity class in the sample spectrum is lost in a hit spectrum of the ASTM file, hit peaks of lower intensities should be regarded as accidentally or erroneously hit, and those hit peaks should be given a penalty rather than a gain. The above equation is based on such a scoring principle, and therefore its use is effective in the reduction of noise spectra in the output list and improving the efficiency of search.

2.5 Mother File and Subfiles

In this study, all of the infrared spectral data file published from the ASTM was used for the test search. The total number of spectra coded therein is 139,800, and hit compounds with higher scores are printed together with their molecular names in the output.

The search time depends on the volume of spectral data file to be searched in the program. In the present search system, the time is roughly proportional to the file volume because most of the search time is spent in the transfer of data from the disk memory to the main memory. For reducing the search time, it will be effective to create several subfiles from the ASTM mother file, to select one of them according to the condition of the input data, and to search only the subfile in the practical search.

In this study, in order to determine how to create those subfiles, we examined the distribution of peak positions coded in the ASTM file. The histogram of the peak wavelength is shown in Fig. 3.

Fig. 3. Histogram of peak wavelengths coded in the ASTM file

By using this result and taking the allowance values of peak wavelengths as described above, fifteen subfiles listed in Table I have been created from the ASTM mother file. In the search program, one of these subfiles is selected out according to which condition is satisfied by the input data of the sample spectrum, and the search is carried out only for the subfile. When no conditions are fulfilled, the mother file is to be searched. It is seen from Table I that the use of the subfile system will reduce the search time at most one fifth as compared with the search time for the ASTM mother file.

The flow sheet for the selection of subfiles in the program is shown in Fig. 4. The condition for the subfiles II, XIII and XV is whether a mandatory peak or peaks exist in the designated wavelength range, and that for the other subfiles is whether no-peak areas include the range.

Table I. Subfiles created from the ASTM mother file

No.	Range(μm)	Y/N ^a	Ratio(%) ^b	No.	Range(μm)	Y/N ^a	Ratio(%) ^b
I	4.6-6.5	N	19.3	IX	9.4-10.0	N	32.3
II	5.8-5.9	Y	29.2	X	10.4-11.1	N	45.0
III	6.0-6.4	N	37.6	XI	11.5-12.5	N	39.0
IV	6.7-6.9	N	25.7	XII	12.9-14.0	N	35.9
V	7.1-7.4	N	27.0	XIII	13.3-13.7	Y	35.5
VI	7.6-8.1	N	24.1	XIV	13.6-15.0	N	43.7
VII	8.3-8.8	N	29.8	XV	14.2-14.7	Y	32.0
VIII	8.8-9.4	N	27.4				

^aY denotes spectra showing at least one peak in the designated wavelength range, and N denotes spectra showing no peak in the range.

^bRatio(%) of the number of corresponding spectra to the total number of spectra in the ASTM mother file.

Fig. 4. Flow sheet for the selection of subfiles

2.7 Test Search

Infrared spectra of 100 compounds were measured with a grating spectrophotometer and used as samples of the test search. The compounds were selected from chemicals to hand, with attention to impartiality in chemical classes. Spectra of liquid compounds were obtained from capillary film, and those of solid compounds were from KBr disk. Peak wavenumbers were calibrated with the polystyrene film, and entered as the input data of the test search.

A FACOM Model M200 computer system was used for the test search. The average search time for the 100 test samples was about 45 seconds on the TSS mode. It is about one third the search time for the mother file, consistent with the figures shown in Table I.

3. Results and Discussion

An example of the result of the test search is shown in Fig. 5. The correct spectra indicated by circles obtain the highest scores, allowing us to identify them easily as the correct answer. The test search using 100 samples proved that the correct spectra always receive the highest scores. This means the advantage of the present search system: correct answers will always exist in spectra of the highest scores.

As pointed out by Erley,⁵⁾ some experience is required for getting a satisfactory result in many of earlier search systems. In those systems, the result of search drastically depends on the selection of peaks to be entered as the input data. If

RESULT OF SEARCH			
①	100	929JA	ACETOPHENONE, P-NITRO-
②	100	5112CA	ACETOPHENONE, P-NITRO-
③	100	F 649KA	ACETOPHENONE, P-NITRO-
④	89	7893FA	ACETOPHENONE, P-NITRO-
5	67	6732CA	BIS-ACETAMIDE, N,NPR-2,7-FLUORENYL-
6	54	5677JA	BENZO//B//THIOPHENE, 6-ACETAMIDO- 2,3-DIBROMO-
7	44	9803FA	ACETIC ACID, P-NITROPHENYL-, METHYL ESTER NYL/-
8	24	10469CA	GLYCINE, N-/2-/P-NITROCARBOBENZOXY- AMIDO/-5-/3-NITROGUANIDO
9	8	14098EA	B-LUMICOLCHICINE
10	8	3923CA	UREA, 1,1-DIPHENYL-3-/B-NAPHTHYL/-

Fig. 5. An example of search result (sample: p-nitroacetophenone)

the number of input peaks is much smaller than the number necessary for a successful search, numerous noise spectra will be printed out and the correct answer will be obscured. Conversely, if too many peaks are entered, the correct answer may be scored too low to be identified easily. Therefore, the number of input peaks for a successful search in many of earlier systems has to be determined critically. In the present search system, on the other hand, such attention is unnecessary, and all peaks even of extremely low peak intensity can be entered as the input data.

References

- 1) D.H.Anderson, G.L.Covert: Anal.Chem.,39,1289(1967).
- 2) D.S.Erley: Anal.Chem.,40,894(1968).
- 3) L.H.Cross, J.Haw, D.J.Shields: "Molecular Spectroscopy", p.189(1968),(The Institute of Petroleum, Adlard, Dorking, U.K.).
- 4) Yu.P.Drobyshev, R.S.Nigmatullin, V.I.Lobanov, I.K.Korobeinicheva, V.S.Bochkarev, V.A.Koptug: Vestn.Akad.Nauk SSSR,40,75(1970).
- 5) D.S.Erley: Appl.Spectrosc.,25,200(1971).
- 6) G.A.Massios: Amer.Lab.,3,55(1971).
- 7) R.W.Sebesta, G.G.Johnson Jr.: Anal.Chem.,44,260(1972).
- 8) S.Shimizu: Proc.Annual Meeting Chem.Soc.Jpn.,2F14(1973).
- 9) S.Kihira: Proc.Tokyo Symp.Appl.Spectrom.,1B11(1973).
- 10) J.Zupan, D.Hadzi, M.Penca: Comp.Chem.,1,71(1976).
- 11) R.N.Jones: Chem.Canada,29,23(1977).
- 12) J.A.deHaseth, H.B.Woodruff, T.L.Isenhour: Appl.Spectrosc.,31,18(1977).
- 13) C.L.Fisk, G.W.A.Milne, S.R.Heller: J.Chromatogr.Sci.,17,441(1979).

Keyword phrases

computer retrieval; infrared spectra; ASTM file search.

(Received July 1, 1981)