CHEMISTRY LETTERS, pp. 521-524, 1982. 

# A NEW LIBRARY SEARCH SYSTEM FOR THE IDENTIFICATION OF MASS SPECTRA

Junko SHINDO,† Akio YASUHARA,* Hiroyasu ITO, and Tsuguo MIZOGUCHI

†Environmental Information Division, National Institute For Environmental Studies,
Division of Chemistry and Physics, National Institute For Environmental Studies, Yatabe, Tsukuba-gun, Ibaraki 305

A new library search system for mass spectra consisting of the pre-search involving seven step filters and the main search modified on the basis of the Probability Based Matching method has been developed.

Gas chromatography-mass spectrometry is one of the most useful techniques for identifying a number of compounds involved in various kinds of environmental samples, since mass spectra are considered compound fingerprints so to speak. Automated library search of mass spectra by computer is designed to shorten search time and, in particular, to identify unknown compounds whose mass spectra have no molecular ion peaks, since search process by manual handling using reference books is irksome and requires much time. Many search systems have already been developed, but only a few are being used routinely with some degree of success.[1-3] Search systems for the retrieval of mass spectra are essentially divided into three groups, that is, the PEAK method,[4] the Biemann method,[5] and the Probability Based Matching (PBM) method.[6,7] The PBM method is the best among the three methods and has two unique features. The first is the probability weighting of masses and abundances and the second is the reverse search proposed by Abramson.[8] An identification of major components is possible even for mixed spectra by the PBM method, but the system requires considerable search time without suitable filters and correct spectra are not often retrieved when the peak intensities of unknown spectra greatly deviate from the reference. Therefore, this system is somewhat inadequate for mass spectra obtained by GC/MS on environmental samples, because of rapidly changing sample concentration and primarily low concentration. This article describes a new library search system applicable to a wide range of use.

The data base for the present system was compiled from the magnetic data tape furnished by EPA/NIH MSDC Data Base which contains the 33,898 mass spectra of different organic compounds. In regard to all peaks (2,187,209 peaks) possessing a relative intensity over 0.5 %, the U and A values were calculated according to the McLafferty method,[7] where U and A are the contribution to the probability of the uniqueness of the m/z value of the peak and the abundance value of the peak

respectively as it appears in the reference spectrum.[9] The 15 peaks of the largest (U + A) values for each spectrum were selected and stored in the data base as reference spectra along with certain information such as the mass number of base peaks, isotopic patterns, rectangular arrays used in the Biemann method, and so forth.

Most of the existing retrieval systems are based mainly on simple matching between reference and unknown spectra and characteristic fragmentation patterns or spectral interpretations have not been considered except interpretative systems[10-12] which do not work effectively yet. Usually, only simple techniques have been applied in most filters to improve search speed by elimination of grossy dissimilar spectra.[13] The filter used in this system is based on several characteristic features in mass-spectral interpretation or fragmentation pattern. The following seven steps are involved in this filter to enhance the reliability of search as well as reduce the search time. (1) The molecular weight for searching ranges from the mass number of the base peak of an unknown spectrum to the mass number of (mass number of maximum mass peak of an unknown spectrum) x 3. (2) Reference spectra must not have peaks of relative intensity exceeding 5 % at a mass number higher than (mass number of the maximum mass peak of an unknown spectrum) + 5. All reference spectra not satisfying this condition are eliminated. (3) Base peaks in reference and unknown spectra must have relative intensities over 50 % at the corresponding mass numbers of unknown and reference spectra, respectively. All reference spectra at variance with this condition are eliminated. (4) Only reference spectra whose rectangular arrays made from the full spectra are similar to that of an unknown spectrum are passed. Namely, over three positions among six ones of the largest sums of intensities in the array must be common to the both spectra. (5) Peaks of even or odd numbered mass with relative intensities over 20 % in reference spectra ( $m/z \geq 60$ ) with even or odd molecular weights respectively should be observed in an unknown spectrum. Reference spectra not satisfying this condition are disregarded. (6) An unknown spectrum must not possess any peak with a relative intensity over 5 % between the mass number of (molecular weight - 4) and the mass number of (molecular weight - 12), based on reference spectra which contain no Cl, Br, S, or Si atoms. All reference spectra at variance with this condition are eliminated. (7) If reference spectra containing Cl, Br, S, or Si atoms show isotopic patterns at the maximum mass area, the unknown spectrum will show the same pattern. Reference spectra satisfying this condition proceed the main search. All steps from the 1st to the 7th step are performed one after another in the case of pure mass spectra. But in the case of mixed spectra, the 4th and the 6th steps and a part of the 3rd step, the checking of reference peak corresponding to the base peak of the unknown spectrum, are not applied. Whether the unknown spectrum was pure or mixed was judged by investigators.

In the main search, the ratio, $\rho$, of the peak intensity of the unknown spectrum to the intensity of reference peak at the same mass number is calculated for each peak of the reference spectrum. If a reference peak with peak number of j is not found in the unknown, $\rho$ is set at zero. The peak with a $\rho$ value less

than the specified threshold ratio, which is 0.15 in this system, is considered to be the unmatched peak, for which bad mark values accumulate. The bad mark values are set to the following values depending on the relative intensity of peak in parentheses: 0 ( < 1 %), 1 (1 % to 10 %), 2 (10 % to 20 %), 3 (20 % to 30 %), 5 ( 30 % to 40 %), 7 (40 % to 50 %), 9 (50 % to 70 %), and 11 ( > 70 %). If the accumulated bad mark values exceed the allowed number which is set to 10 in this system, this reference spectrum is eliminated as a dissimilar one. Then a new parameter, KS, is defined as a measure of confidence for search in equation (1), where k equals the number of matched peaks.

$$KS = \sum_{j}^{k} ( U_j + A_j ) / \sum_{j}^{15} ( U_j + A_j ) \quad (1)$$

If all the reference peaks are contained in the unknown spectrum with reasonable intensities, KS has a maximum of 1.0.

In order to estimate the proportion of contamination from other mass spectra but not from the retrieved spectrum, a new parameter, PC, is defined in equation (2) as,

$$PC = ( \sum_{j}^{h} I_j^U + \sum_{j}^{k'} ( I_j^U - \alpha_j I_j^R \rho_{min} ) ) / \sum_{j}^{h+k} I_j^U \quad (2)$$

The product of the intensity ( $I_j^R$ ) of each matched reference peak and $\rho_{min}$ , that is the smallest $\rho$ value among the matched peaks, is calculated. This product ( $I_j^R \rho_{min}$ ) is an expected minimum intensity for the reference peak with the number of j in the unknown spectrum and the allowable maximum intensity for the same peak is set $\alpha_j I_j^R \rho_{min}$ , where $\alpha_j$ is a factor empirically determined and varies from 2.0 to 4.0 according to the intensity of reference peak. If the actual intensity ( $I_j^U$ ) of the unknown peak is greater than $\alpha_j I_j^R \rho_{min}$ , it is considered to be contaminated. The PC value which is the proportion of contamination is calculated for the ten peaks, having the largest (U + A) values in the unknown spectrum, and all matched peaks according to equation (2), where h is the number of peaks not found in the reference spectrum for the ten peaks in the unknown spectrum, k' is the number of matched peaks whose intensities are greater than $\alpha_j I_j^R \rho_{min}$ , and k in the denominator equals the number of matched peaks. Another measure of confidence, KD, is introduced by KS and PC in equation (3).

$$KD = KS ( 1 - PC ) \quad (3)$$

If the estimated KS or KD values are greater than the specified threshold values, KS, KD, and PC values are stored with the corresponding compound name, molecular formula, and molecular weight. These stored data are printed after completion of the search. The mass spectra of 80 different organic compounds obtained at an ionizing energy of 75 eV with a JEOL Model JMS-D 100 mass spectrometer connected to a JGC-20K gas chromatograph and a JMA-2000 mass data analysis system were tested by this library search system.

Search by this system was compared with the common PBM method using the same filters by the same computer system (HITAC M-180) and the same data base in the National Institute For Environmental Studies. In addition to the K values,[6] the K % value, proposed by Kato et al.,[14] being the ratio of the estimated K value to

the maximum K value computed for the spectrum perfectly matched with the reference, was used as a measure.

Reference spectra retrieved with the present system were ordered by decreasing KS value. If two reference spectra have the same KS value, the reference spectrum with larger KD is superior to another one. Retrieved spectra by the PBM method were also sorted in the same manner by using K % and K values. If a correct answer of an unknown spectra is involved in the 4 reference spectra at the top of this ordered list, it is considered that the library search has been successful. Ratio of successful search was 69 % in this system and is very high compared to the value of 21 % obtained with the PBM method. Twenty unknown spectra were searched within a maximum period of 3 minutes by HITAC M-180 computer system. The filter used in the pre-search of this system was found to be sufficiently effective. Out of 33,898 reference spectra, for example, the following number of spectra passed the respective steps from the 1st concerning molecular weight range to the 7th concerning isotopic pattern, in the search for butyl benzoate: 1st step, 21436; 2nd step, 7856; 3rd step, 48; 4th step, 38; 5th step, 10; 6th step, 10; and 7th step, 8. It is concluded that this new library search system is very valuable in view of the results of successful search and should be very practical for rapid search.

We should like to express our appreciation to Dr. Yoshiko Kato and his colleagues for their useful comments.

References

1) F.W.McLafferty and R.Venkataraghavan, J.Chromatogr.Sci., 17, 24(1979).
2) D.Henneberg, Adv.Mass Spectrom., 8B, 1511(1980).
3) D.P.Martinsen, Appl.Spectrosc., 35, 255(1981).
4) S.R.Heller, Anal.Chem., 44, 1951(1972).
5) H.S.Hertz, R.A.Hites, and K.Biemann, Anal.Chem., 43, 681(1971).
6) F.W.McLafferty, R.H.Hertel, and R.D.Villwock, Org.Mass Spectrom., 9, 690 (1974).
7) G.M.Pesyna, R.Venkataraghavan, H.E.Dayringer, and F.W.McLafferty, Anal.Chem., 47, 1362(1976).
8) F.P.Abramson, Anal.Chem., 47, 45(1975).
9) G.M.Pesyna, F.W.McLafferty, R.Venkataraghavan, and H.E.Dayringer, Anal.Chem., 47, 1161(1975).
10) J.R.Chapman, "Computers in Mass Spectrometry," Academic Press, London(1978), p.187.
11) R.M.Hilmer and J.W.Taylor, Anal.Chem., 51, 1361(1979).
12) L.W.McKeen and J.W.Taylor, Anal.Chem., 51, 1368(1979).
13) J.R.Chapman, "Computers in Mass Spectrometry," Academic Press, London(1978), p.139.
14) J.Shishido, Y.Kato, Y.Takubo, T.Yamamoto, M.Fujii, T.Yamashita, M.Kato, and M.Kondo, Seikatsu Eisei, 23, 157(1979).

(Received February 4, 1982)