Vol. 95 No. 1/2 September 15, 1977
(Computer Techniques and Optimization, Vol. 1, No. 1/2)
ISSN 0378-4304

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry
Revue internationale consacrée à tous les domaines de la chimie analytique
Internationale Zeitschrift für alle Gebiete der analytischen Chemie

PUBLICATION SCHEDULE FOR 1977 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	88/1	88/2	89/1	89/2	90	91/1	91/2	92/1	92/2	93	94/1	94/2
Section on Computer Techniques and Optimization									95/ 1+2			95/3+4

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with every aspect of modern chemical analysis, both fundamental and applied. The section on *Computer Techniques and Optimization* is devoted to new developments in chemical analysis by the application of computer techniques and by interdisciplinary approaches, including statistics, systems theory and operation research.

Submission of Papers. Manuscripts (three copies) should be submitted to:
for *Analytica Chimica Acta*: Dr. A.M.G. Macdonald, Department of Chemistry, The University, P.O. Box 363, Birmingham B15 2TT, England.
for the section on *Computer Techniques and Optimization*: Dr. J.T. Clerc, Laboratorium für Organische Chemie, Swiss Federal Institute of Technology, Universitätstrasse 16, CH-8092 Zürich, Switzerland.

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 93, p. 379 for detailed information. Reprints of this information are available from the Editors or from: Elsevier Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DE (Great Britain).

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An order form containing price quotations will be sent to the authors together with the proofs of their article.

Advertisements. Advertisement rates are available from the publisher.

Subscriptions. Subscriptions should be sent to: Elsevier Scientific Publishing Company, P.O. Box 211, Amsterdam, The Netherlands. The section on *Computer Techniques and Optimization* can be subscribed to separately.

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 8 volumes in 1977. The subscription for 1977 (Vols. 88–95) is Dfl. 920.00 plus Dfl. 112.00 (postage) (Total approx. US $ 420.95). The subscription for the *Computer Techniques and Optimization* section only (Vol. 95) is Dfl. 115.00 plus Dfl. 14.00 (postage) (Total approx. US $ 52.75). Journals are sent automatically by air mail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–87) are available at Dfl. 115.- (plus postage).

Claims for issues not received should be made within three months of publication of the issue, otherwise they cannot be honoured free of charge.

Printed in The Netherlands

ANALYTICA CHIMICA ACTA COMPUTER TECHNIQUES AND OPTIMIZATION

You are invited to subscribe

Please use the cards below

ANALYTICA CHIMICA ACTA **ORDER FORM**
COMPUTER TECHNIQUES AND OPTIMIZATION
1977: VOLUME 1 In 4 ISSUES (Volume 95 of Analytica Chimica Acta)
The new section may be subscribed to separately. Subscribers to Analytica Chimica Acta will receive it automatically.

☐ Please enter a subscription starting with volume 1 (1977)
at $ 52.75/Dfl. 129.00 including postage
☐ Please enter a subscription starting with volume 2 (1978)
at $ 57.25/Dfl. 140.00 including postage

Orders from individuals must be accompanied by a remittance.

I enclose | my personal cheque | bank draft | UNESCO coupons |

☐ Please send a free specimen copy

Journals are automatically sent by air to the U.S.A. and Canada, at no extra cost and to Japan, Australia and New Zealand with a small additional postal charge.

Signature Date _____

Name _____

Address _____

_____

_____ Postal Code _____

The Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

ANALYTICA CHIMICA ACTA **ORDER FORM**
COMPUTER TECHNIQUES AND OPTIMIZATION
1977: VOLUME 1 In 4 ISSUES (Volume 95 of Analytica Chimica Acta)
The new section may be subscribed to separately. Subscribers to Analytica Chimica Acta will receive it automatically.

☐ Please enter a subscription starting with volume 1 (1977)
at $ 52.75/Dfl. 129.00 including postage
☐ Please enter a subscription starting with volume 2 (1978)
at $ 57.25/Dfl. 140.00 including postage

Orders from individuals must be accompanied by a remittance.

I enclose | my personal cheque | bank draft | UNESCO coupons |

☐ Please send a free specimen copy

Journals are automatically sent by air to the U.S.A. and Canada, at no extra cost and to Japan, Australia and New Zealand with a small additional postal charge.

Signature Date _____

Name _____

Address _____

_____

_____ Postal Code _____

Postcard

Place stamp here

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

P.O. Box 211

AMSTERDAM
The Netherlands

Postcard

Place stamp here

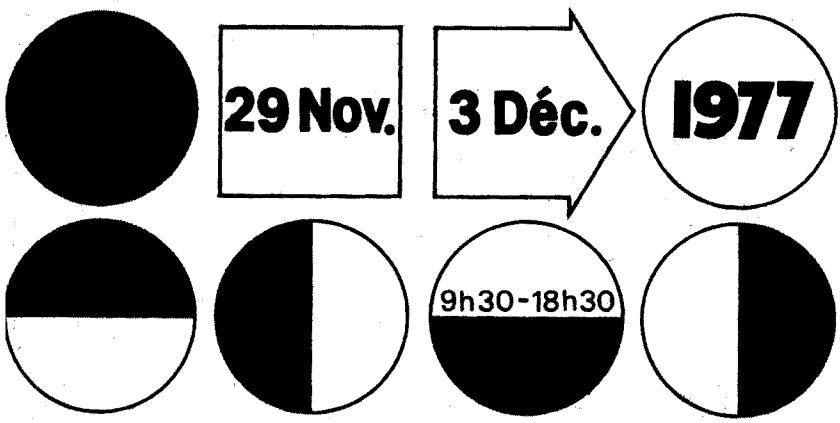
ELSEVIER SCIENTIFIC PUBLISHING COMPANY

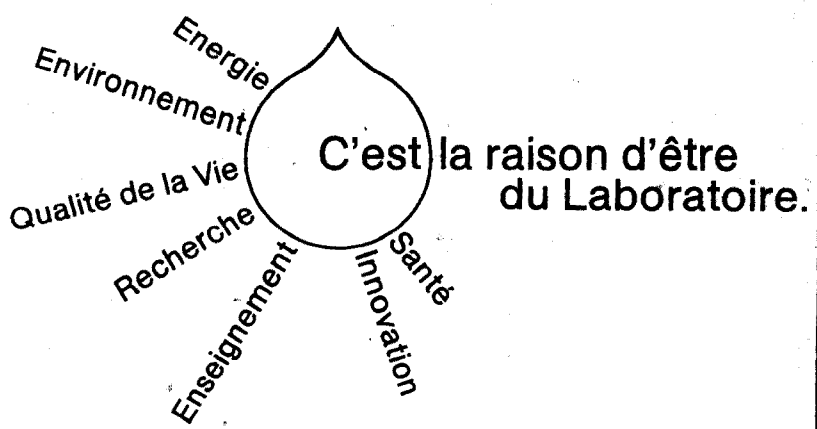
P.O. Box 211

AMSTERDAM
The Netherlands

salon du laboratoire 1977

Paris, porte de Versailles





La 67e Exposition de Physique a lieu conjointement avec le Salon du Laboratoire. L'entrée des deux expositions est commune.

Le Congrès de Chimie Analytique - 33e Congrès du G.A.M.S. - se tiendra aux mêmes lieux et dates que ce salon.

Salon organisé par l'Association pour le Salon du Laboratoire régie par la loi de 1901
12, rue Chabanais - 75002 PARIS - France - Tél. 742.79.00

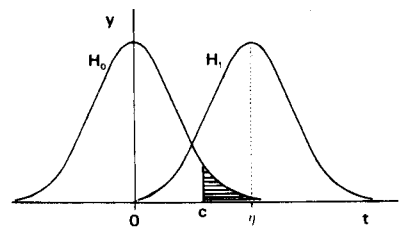
9

Verdere inlichtingen, folder en voordelig reisarrangement worden gaarne verstrekt door de
STICHTING TER BEVORDERING VAN DE FRANSE VAKBEURZEN
Prins Hendrikkade 20-21 - Amsterdam-C. - Telefoon (020) 248670 en 239204

Statistical Treatment of Experimental Data

J. R. GREEN, *Lecturer in Computational and Statistical Science,* University of Liverpool, *and* D. MARGERISON, *Senior Lecturer in Inorganic, Physical and Industrial Chemistry, University of Liverpool.*

This book is intended for researchers wishing to analyse experimental data using statistical methods. Statistical concepts and methods which may be employed, are explained, and the ideas and reasoning behind statistical methodology clarified. Formal results are illustrated by many numerical worked examples mainly taken from the laboratory. Concepts, practical methodology, and worked examples are integrated in the text.



Consideration is given in this work to a large number of practical topics which are often omitted from standard texts. These include: obtaining an approximate confidence interval for a function of some unknown parameters; testing for outliers, stabilization of heterogeneous variances, and significant differences between means; estimation of parameters after performing tests; deciding what numbers of significant figures to quote for sample means and variances; straight-line and polynomial regression, through the origin or not, using weighted points, and testing the homogeneity of a set of such lines or curves.

The many examples provided throughout the text will serve as models for the various problems encountered by the readers when employing statistical methods to treat experimental data. Neither a strong mathematical background nor a prior knowledge of probability or statistics is required in order to make use of this work.

In addition to research workers in universities and industry, the book will be of use for first-year students of statistics, and will be especially suitable as the basis of a graduate course in experimental sciences.

Sept. 1977 xiv + 382 pages US $34.95/Dfl. 85.00 ISBN 0-444-41615-3

ELSEVIER

P.O. Box 211, Amsterdam
The Netherlands
52 Vanderbilt Ave
New York, N.Y. 10017

The Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

7046

ANALYTICA CHIMICA ACTA
VOL. 95 (1977)
(Computer Techniques and Optimization, Vol. 1)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

VOL. 1 1977

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

Anal. Chim. Acta, Vol. 95 (1977)

PRINTED IN THE NETHERLANDS

Analytica Chimica Acta, 95 (1977) 1
Computer Techniques and Optimization

EDITORIAL

The steady progress in the development of modern electronic devices in recent years has greatly affected the area of analytical instrumentation. However, the introduction of computer techniques has not only influenced instrumental design but has led to new sophisticated methods of measurement and to the modification of older methods that can be applied routinely in the laboratory. New methods of interpreting analytical measurements, e.g. the use of computer-searchable data collections, have become feasible; as a result, the reliability and the amount of readily accessible analytical information have already increased greatly.

In acknowledgement of these developments, it has been decided to devote a special section of *Analytica Chimica Acta* to the various applications of computer techniques in analytical chemistry. It should be emphasized that this section is not regarded as a new journal, but merely as a specialized section of a long-established international journal. It is hoped that this will enable many papers relating to computer techniques and methods of optimization to be concentrated in one section of one journal, thus saving time-consuming literature searches.

The editors of this new section of *Analytica Chimica Acta* will be glad to consider for publication all manuscripts dealing with computerized techniques and with any aspects of automation — as opposed to mechanization — in analytical chemistry. It is hoped that the combined efforts of a new team of editors and editorial advisers — all with established reputations within this new and specialized field — will lead to the rapid establishment and international acceptance of this new section as a major centre of communication in this developing field.

J. T. Clerc
E. Ziegler

Analytica Chimica Acta, 95 (1977) 3—12
Computer Techniques and Optimization

HIERARCHICAL PREPROCESSING OF INFRARED DATA FILES

M. PENCA, J. ZUPAN* and D. HADŽI

Chemical Institute "Boris Kidrič", and Faculty of Natural Sciences and Technology, University of Ljubljana, Ljubljana (Yugoslavia)

(Received 29th November 1976)

SUMMARY

A preprocessing method for the use of infrared data files in structural elucidation is discussed. Different hierarchical trees, as well as multi-category predictions based on average infrared spectra of the appropriate classes are prepared and inspected for predictive ability.

The first stage in the application of computers in speeding up identification processes is based on matching the spectrum of an unknown compound against the spectra of known compounds stored in a collection. Initially, the matching was done by mechanical sorting machines, and this relieved spectroscopists from the most time-consuming and dull part of the task. The advent of digital computers shortened the process by several orders of magnitude and thus made practicable searches over libraries containing ca. 10^5 spectra in a matter of minutes.

However, searching such large numbers of spectra is not only uneconomical in time even with the best program and fastest computer but also wastes a good deal of the information contained in the spectral library as well as the logic capacity of the computer. This could be true even for ideal matching of the unknown and known spectra. If this were so, the search for a spectrum not in the library, would yield the rather trivial result that the unknown is none of the 10^5 compounds listed. However, the search programs allow a certain latitude both in the presence and the frequency of peaks; this is necessary because of unavoidable recording and coding errors. Accordingly, several spectra are usually obtained from the search; the more the limits are relaxed, the greater the number of spectra.

The non-ideality of the search system has led to the development of rating algorithms [1—4] for the similarities of both spectra and structural features. However, this does not bring any closer the second stage in the application of computers in spectra-based analysis. In this stage the information stored in the spectra file should be used to obtain information regarding the identity of the compound even if the matching spectrum is not contained in the file. For this purpose, another strategy is required, and the file must be organized in a different structure. In this context, the term

"structure of the file" no longer refers to how the data are arranged within the records but how the records are linked and clustered or addressable in the file.

Because of the organization of the file, a search over the whole file then becomes unnecessary. This may well be the most important effect, at least initially. These features are intimately connected. The problems of how to establish the best structural features and how to use them in the further process is very hard and complex. After the features relevant to the spectroscopic data (peak positions, intensities, peak shapes, multiplicities, presence or absence of some peaks or groups of peaks, etc.) have been obtained, the construction of sound, well-defined hierarchical trees seems to be a promising way to proceed. The decision points of such a tree should differentiate between structural features; such a tree should not be rigid but flexible with overlapping decisions.

In the present work only a small part of this huge problem was attacked: some kind of "average" infrared spectra for different kinds of hierarchical classifications were tested to discover the one giving the best prediction for the structure of 200 test spectra.

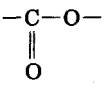
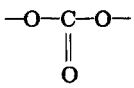
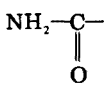
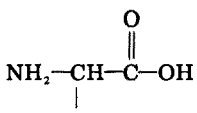
DATA BASE

The data base for the present work was the Wyandotte-ASTM Infrared Spectra File [5] containing about 93,000 infrared spectra. The spectra of all compounds with a coded >C=O fragment (further referred to as V-compounds, as in the WLN [6] symbol for the >C=O fragment) were picked out from the entire collection (about 2500) by means of the retrieval system ZAPAH 2 [3, 4]. From these 2500 spectra, 781 carefully selected spectra were taken as a basic set. The test was mainly done for miscoding and on the basis that each type of V-compound was represented by approximately the same number of compounds. The selected 781 V-compounds were divided into 7 classes, as shown in Table 1. From the rest of the collection a corresponding total of about 500 compounds (with the V fragment absent) was chosen randomly.

Because of the known shortcomings of the Wyandotte collection [4] (the lack of intensity information and large tolerance regions for miscoded peak positions), each peak in the selected spectra represented by 140 bits (2.0—15.0 μm) was replaced by a Gaussian peak with a range of ± 0.2 μm (5 bits) with the maximum intensity of 1. In the case of overlapping bands, the intensities were normalized to the strongest one. Thus the spectrum (140 bits, 1 or 0) of compound i belonging to class j was replaced by a 141-dimensional vector $\mathbf{X}_i^j$ containing intensity data on components 1 to 140 with information on class j (or type of fragment) in the 141st position. Each component $\bar{x}_k^j$ of the "average" spectrum $\mathbf{X}^j$ for the class j was simply obtained as the average over the corresponding components of all vectors

TABLE 1

Classes of V-compounds
(Classes 2 and 4 were not further divided into 2 and 3 classes, respectively, because they cannot be distinguished by the ASTM coding system [5].)

Class	Fragment	WLN	Name	No.
1		VQ	Carboxylic acids	93
2		VO	Esters, lactones, anhydrides	175
3		VH	Aldehydes	105
4		V	Ketones, Acid halides	120
5		OVO	Carbonates	94
6		ZV	Amides	95
7		ZYVQ	Amino acids	99
			Total	781

$\mathbf{X}_i^j$ belonging to the same class j

$$\bar{\mathbf{x}}_k^j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{x}_{i,k}^j \qquad (1)$$

where N_j is the number of spectra in class j. If during the work some "average" spectra for the group containing compounds with different fragments were requested, the summation in eqn. (1) was extended over all spectra belonging to the desired classes. Figure 1 shows the "average" spectra of the 7 "pure" classes listed in Table 1.

As the test set, 213 spectra from the Aldrich Catalog [7] were selected. The structure of this set is shown in Table 2. To be comparable with the training set, the test spectra were digitized in the same way as the library file (140 bits per spectrum) and then extended into the 141 dimensional vector as described above. This is an immense waste of information, but some kind of standardization of spectra with respect to the intensities is necessary, and this method seems to be quite suitable, especially when spectra from different sources must be tested.

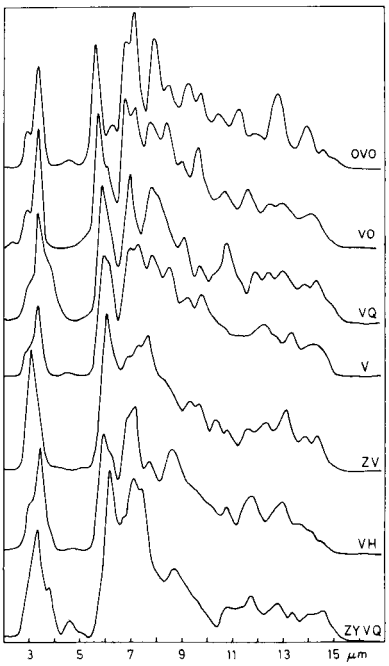


Fig. 1. Average spectra of 7 V-compounds used for the 7-class multi-decision.

THE DECISION

Each single decision in the hierarchical tree was always made on the basis of comparison between two average spectra: one of a given group k (containing compounds of one or more types j) and the other of all compounds not belonging to the group k but not yet rejected by some previous decision. For each decision point k, the so-called difference vector ΔX^k was calculated by subtracting both average vectors: $\Delta X^k = \bar{X}^k - \bar{X}^{\text{not } k}$. An example of such a difference vector is shown in Fig. 2.

It is clear that the difference vector is very suitable for the location of important regions in the spectrum for this particular decision: the greater the difference the more significant is the region and vice versa. For each difference vector, a reasonable limit below which the comparison is useless (or even error accumulating) must be determined. It was found that there is no fixed way of choosing this limit; it was estimated simply by a trial-and-error procedure for each particular case.

Finally the decision was made by restricting the comparison (subtraction) of the unknown spectrum with both average spectra in those regions where the difference spectrum for this decision was above the estimated tolerance limit. The smaller difference from both average spectra indicates the category to which the unknown compound belongs. Table 3 shows the dependence on the tolerance limit of the recognition ability for some decision points.

TABLE 2

The structure of the 213 test compounds

Fragment	Class	Compound	Aliphatic	Aromatic
	0	Hydrocarbons	11	4
	0	Halogenated hydrocarbons	7	4
	0	Alcohols	6	4
	0	Ethers, acetals, epoxides	4	4
	0	Mercaptans, sulfides	3	2
	0	Amines	4	10
	0	Nitro-, nitroso- compounds	4	4
	0	Acid salts	5	1
	0	Nitriles and cumulative double bonds	4	2
	0	Sulfur-oxygen compounds	4	3
	0	Phosphorus compounds	2	3
	0	5-Membered heterocycles	—	7
	0	6-Membered heterocycles	—	7
	0	Heterocyclic *N*-oxides	—	2
	0	Oximes	2	—
QV	1	Carboxylic acids	5	7
VO	2	Esters, lactones, anhydrides	17	10
VH	3	Aldehydes	5	5
V	4	Ketones, acid halides	10	14
OVO	5	Carbonates	2	2
ZV	6	Amides	6	7
ZYVQ	7	Amino acids	5	5
		Totals	106 +	107 = 213

TABLE 3

The dependence of recognition ability on the tolerance limit
(The tolerance limit is given in units of the difference vector, i.e. difference in intensities of corresponding average spectra, see Fig. 2.)

Decision		Tolerance limit (%)			
		5	10	15	20
Carbonates	: yes or no	74.9	75.0	79.0	82.5
Amino acids	: yes or no	74.2	73.9	71.8	70.7
Aldehydes	: yes or no	69.5	73.0	78.5	75.2

CONSTRUCTION OF HIERARCHICAL TREES

In a search for the best solution for the present choice of fragments, several different trees were tested for their predictive ability; the best two are shown in Figs. 3 and 4. For the sake of comparison a 7-class multi-prediction was also tested and will be discussed later.

The first decision in the present study is dictated in advance by the choice of fragments and not by inspecting the whole file in some statistical way although this step will not be necessary in the real, completed information

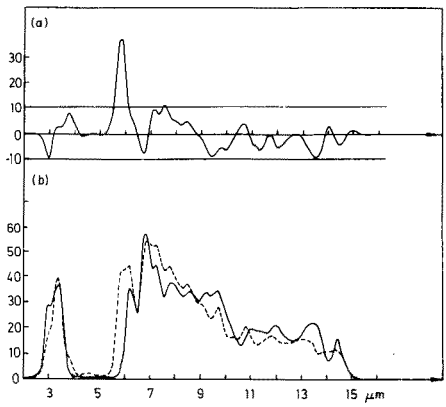


Fig. 2. Pair of average spectra representing V- and non V-compounds (below) together with the difference vector (above). The intensities of the average spectra are calculated from eqn. (1).

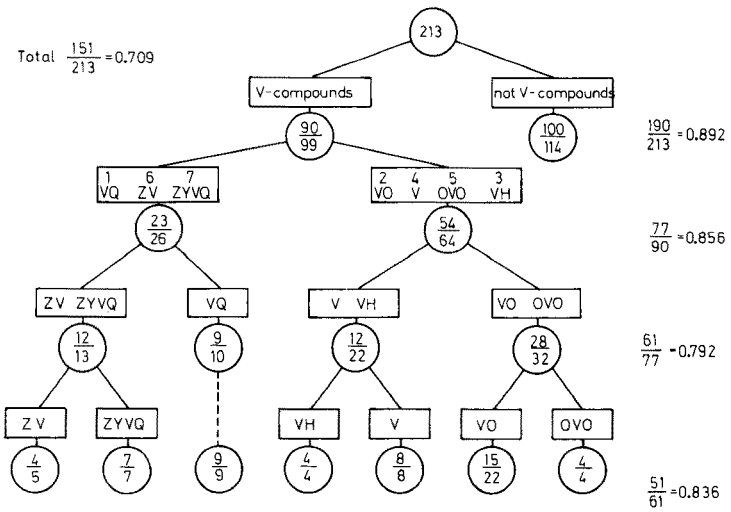


Fig. 3. The hierarchical tree with the best recognition and predictive ability. The binary decisions in this tree are made with average vectors shown in Figs. 5 and 6.

system. The selection of the decision as to whether the compound belongs to the V-compound or not is based on the general experience that the decision is very easy.

First, the average vectors of V- and non V-compounds, together with the difference vector, were obtained (Fig. 2) and tested. The recognition and the predictive abilities were relatively poor (about 75%). The reason for this failure can be clearly seen from the difference vector (Fig. 2a) between both vectors representing V- and non V-compounds (Fig. 2b): the significant part, with the lower limit of 10%, is restricted to the narrow region 5.6—6.3 μm. This region covers the most significant part of the spectra representing V-compounds but is not sufficiently selective in either the positive or negative sense, as many non V-compounds also have characteristic peaks

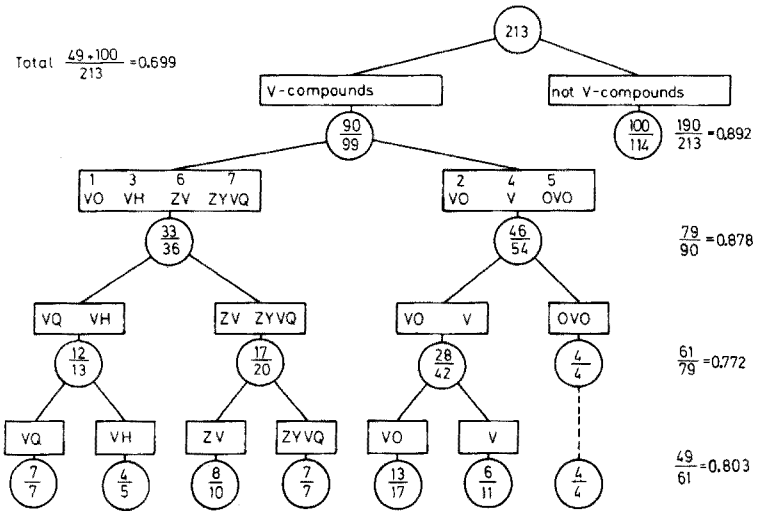


Fig. 4. The hierarchical tree with the second best recognition and predictive ability. This tree gives better prediction than the first one up to the last level.

in this region (all aromatics, for instance). Moreover, carbonates and amino acids which are also V-compounds have peaks very close to 5.6 and 6.3 μm, respectively: wrong decisions therefore occur very easily.

Some other possibilities were therefore tried; finally, a decision based on 7 pairs of different average vectors was selected. Each pair of average vectors represents the class of compounds having the j-th fragment versus compounds containing all other non-j fragments and all non V-compounds. This quasi-multi prediction was used for sorting the V- from the non V-compounds only, regardless of the particular decision. Only if the unknown compound was not classified in any of the 7 classes was it rejected as a V-compound, otherwise it was treated further as a V-compound. The predictive ability for this decision was about 75%.

After the first decision has been made, two alternative ways of constructing the structure elucidation process remain: either a one-level multi-decision or binary decisions sequentially ordered in a hierarchical tree. The multi-decision based now on 7 average vectors of V-compounds (Fig. 1) has a predictive ability of 61.1% (Table 4).

After the spectrum of the unknown test compound has been subtracted from all 7 average vectors, the smallest difference shows the category to which it belongs. However, one serious shortcoming of all multi-category decisions should not be overlooked. Once the decision has been made, the link with the neighbours or the similar compounds is lost; this becomes worse as the multiplicity of decision increases.

The second choice, i.e. the binary decision tree, seems better because of its simplicity and familiarity to the spectroscopist. The group of 7 classes could be divided in 35, 21, or 7 ways into two groups containing each of

TABLE 4

The predictive ability of 7-class multi-decision tested on spectra

Class	Fragment	No. of compounds	No. of correct classifications
1	VQ	12	6
2	VO	28	25
3	VH	10	8
4	V	22	5
5	OVO	4	2
6	ZV	8	6
7	ZYVQ	6	3
Total		90	55

3 and 4, 2 and 5, or 1 and 6 classes, respectively. The recognition ability was tested for all 63 possibilities. The results could be improved more easily if the source group was divided into approximately two equal subgroups rather than into two unequal ones. The best results are shown in Table 5. The best recognition ability of all 3/4 groups exceeded the best for 2/5 groups, while all 1/6 groups gave recognition of below 70%.

The detailed optimization of the decision tree with respect to the final recognition ability was made on the 7 classes of V-compounds shown in Table 4. In spite of the best recognition of the highest rated 3/4 group, i.e. separation into class (1, 3, 6, 7) versus (2, 4, 5), the final results were better for the separation (1, 6, 7) vs. (2, 3, 4, 5), as was also proved on the test set. The optimal trees for both best decisions are presented in Figs. 3 and 4, while all average spectra used for the whole decision tree, together with the appropriate difference vectors, are shown in Figs. 5 and 6.

TABLE 5

Recognition ability for different groupings of V-compounds

Type	Class							Recognition (%)
	VQ 1	VO 2	VH 3	V 4	OVO 5	ZV 6	ZYVQ 7	
3/4	1	−1	1	−1	−1	1	1	84.9
	1	−1	−1	−1	−1	1	1	81.4
	1	−1	1	−1	−1	−1	1	78.1
	1	−1	−1	1	−1	1	1	77.9
all other 3/4 groups give results poorer than 75%								
2/5	1	−1	1	−1	−1	−1	−1	79.1
	1	−1	−1	−1	−1	−1	+1	76.3
all other 2/5 groups give results poorer than 75% all 1/6 groups give results poorer than 70%								

CONCLUSION

In the present study, the training set of spectra plays a different role from that in the usual types of learning machine procedures. The training set, or rather basic set, serves only to form the average vectors and is used only twice, for a given binary decision, to form both representative vectors. All predictive abilities in the present study are slightly greater than the recognition abilities, probably because many more vectors are used in the basic set than for testing, hence a further, small decrease in the predictive ability could be expected. From the source of the basic set [5], some hidden errors are also still present; all spectra used for a test were checked and digitized by the authors.

In all trials, the largest source of wrong decisions was the group of ketones and/or esters. By inspecting the basic as well as the test spectra, it was found that these wrong decisions were caused mainly by the anhydride group (VOV fragment) which was grouped with the ester group as they cannot be distinguished from each other by the coding notation of ASTM [5]. With ketones, the aromatic or aliphatic compounds were frequently decided wrongly.

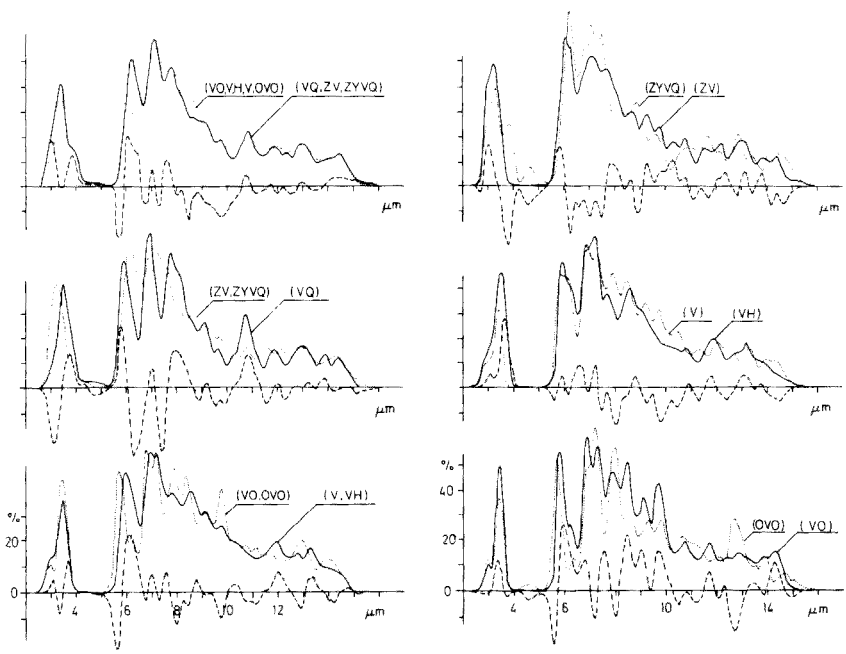


Fig. 5. (Left) Average and difference vectors for the first three decisions in the best hierarchical tree. The average vectors are represented by full (category + 1) and by dotted lines (category −1); the difference vector is dashed.

Fig. 6. (Right) Average and difference vectors for the last decision level in the best tree shown in Fig. 3.

The proper place for the hierarchical tree in a complex retrieval system is shown in Fig. 7. Up to the n-th level, the carefully selected and constructed tree shortens the search procedure by a factor of 2^n, compared with the usual procedures. The second advantage is, of course, the positive information about the structure of the unknown compound if the subsequent search in the short 2^n th part of the whole file fails to yield an adequate or reliable result, i.e. if the matching spectrum is not contained in the reference collection.

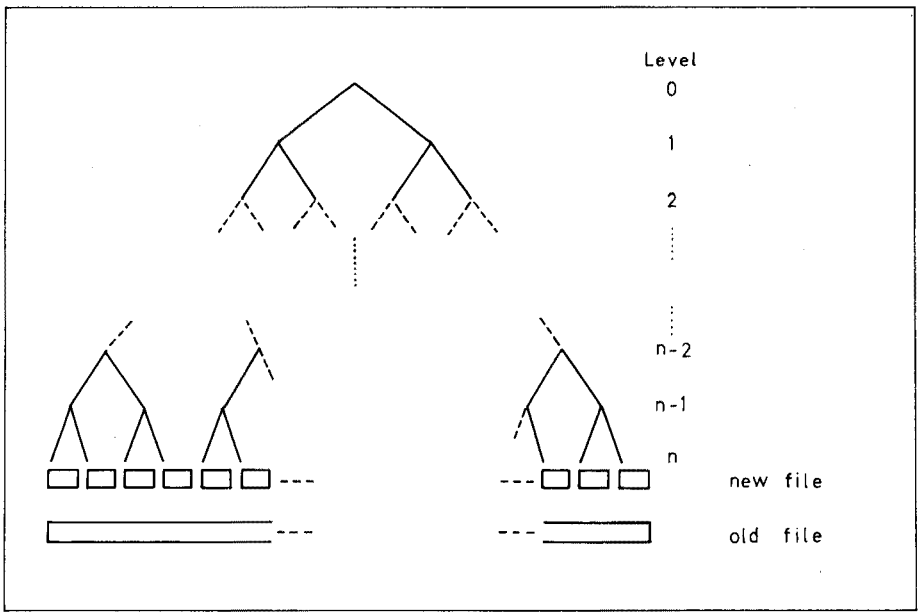


Fig. 7. The hierarchical decision tree selected before the search is applied on a small part of the file. The old library file has to be divided in advance into small parts according to the tree.

We wish to acknowledge financial support from the Research Community of Slovenia and the Pharmaceutical Factory KRKA.

REFERENCES

1 D. S. Erley, in D. Hadži (Ed.), Computers in Chemical Research and Education, Elsevier, Amsterdam, 1973, Vol. I, p. 2/97.
2 P. R. Naegeli and J. T. Clerc, Anal. Chem., 45 (1974) 739 A.
3 J. Zupan, D. Hadži and M. Penca, Comp. Chem., 1 (1976) 71.
4 J. Zupan and D. Hadži, III International Conference on Computers in Chemical Research, Education, and Technology, Caracas, Venezuela, (1976).
5 Codes and Instructions for Wyandotte-ASTM, 1916 Race St., Philadelphia, Pa., (1964).
6 E. G. Smith, The Wiswesser Line-Formula Chemical Notation, McGraw-Hill, New York, (1968).
7 The Aldrich Library of Infrared Spectra, Aldrich Chemical Co. Inc., 1970.

Analytica Chimica Acta, 95 (1977) 13—23
Computer Techniques and Optimization

COMPUTER-ASSISTED INTERPRETATION OF INFRARED SPECTRA

HUGH B. WOODRUFF and MORTON E. MUNK*
Department of Chemistry, Arizona State University, Tempe, Arizona 85281 (U.S.A.)
(Received 2nd March 1977)

SUMMARY

Pattern recognition and artificial intelligence programming techniques for the interpretation of infrared spectra are compared in an effort to determine the best technique for assisting the solution of actual structural elucidation problems. For several reasons, artificial intelligence is the method of choice. When information pertaining to a large number of classes is required, an excessively large training set would be needed with pattern recognition procedures. If the program makes a mistake, it must be alterable so that a similar error will not occur again. Artificial intelligence programs are amenable to this correction procedure.

Previous work [1] has resulted in the development of an artificial intelligence program designed to assist in interpreting infrared spectral data. This program is specifically designed to interface to a package of computer programs that assist in solving actual structural elucidation problems. The results from the infrared interpreter are used as direct input for a structure generator program. Several stipulations are placed on the interpreter to enhance its utility in the structure elucidation package. This paper discusses those stipulations and demonstrates why the artificial intelligence approach is the method of choice for this specific situation. During the process of this demonstration, the artificial intelligence program is compared to several pattern recognition procedures.

All programs were coded in FORTRAN IV and run on the Arizona State University Univac 1110 computer.

Background

Much of the emphasis with computerized structure elucidation schemes has been placed on the development of rapid and efficient structure generator programs [2—6]. These programs combine polyatomic fragments and any residual atoms from the molecular formula to obtain plausible candidate molecules. It is necessary to supply a sufficient number of fragments to prevent the generation of an unwieldy number of candidate structures, and to assist with this task, an interactive artificial intelligence infrared interpretation program has been designed [1]. If the program is to be of general value, it must be capable of interpreting the relatively complex spectra of

compounds found in nature. Additionally, decisions must be made concerning the presence or absence of a large number of functional groups. Finally, it is unnecessary for the program to attempt a definitive interpretation of the spectrum and pass the results directly to the structure generator. Rather, it is important that the program should make a logical interpretation of the infrared spectrum and immediately present the information. The program output can then be reviewed along with any supplementary data available, so that a final decision can be made on those fragments to be used as input for the structure generator. By design, the chemist plays an integral role in the structural elucidation process. The most important function of the spectral interpreter is to provide ideas on how all the information available can best be utilized.

In an effort to keep the present investigation at a manageable size, only one classification problem has been studied in detail; namely, the programs must determine whether each compound is an acid or not. This distinction is usually relatively simple to make from infrared spectral data.

PATTERN RECOGNITION APPROACHES

The results of a variety of pattern recognition procedures on classifying infrared data have been reported [7—10]. Some of this work has used data collected by the American Society for Testing and Materials (ASTM). This ASTM collection contains binary (peak/no peak) data, thus eliminating intensity information. It has been demonstrated for both infrared data and mass spectral data that binary data can be classified correctly with about the same frequency as data that include intensities [11—13]. Two different pattern recognition approaches are used in this study: distance from the mean [13] and similarity measures (nearest neighbor) [14].

The approach involving distance from the mean requires a training set, in this case divided into acid and non-acid classes. A weight vector for class i, $\mathbf{W}_i$, is obtained by approximating the class conditional probabilities for the class i spectra (p_{ji}) for each dimension j

$$p_{ji} = \sum_{n=1}^{m_i} x_j/m_i \tag{1}$$

where $x_j = 0$ or 1 and m_i = number of spectra in class i.

In the case of the ASTM infrared data, 139 dimensions are employed, corresponding to the 139 0.1-μm intervals between 2.0 and 15.9 μm. The discriminant function is obtained by measuring the distance of a spectrum from both the class and non-class weight vectors

$$D_i(\mathbf{X}) = [(\mathbf{X} - \mathbf{W}_i) \cdot (\mathbf{X} - \mathbf{W}_i)]^{\frac{1}{2}} \tag{2}$$

The spectrum is predicted to belong to the class for which its weight vector is closest to the spectrum [15].

Classification based on a similarity measure still requires a data set of known spectra, but this data set is not actually a training set as no weight vectors are calculated. The unknown spectrum is compared with each data set member and predicted to belong to the same class as the member most similar to it. (Alternatively, a majority vote may be taken among the k most similar data set members and the unknown predicted to belong to the majority class.) Two similarity measures — nearest neighbor and Tanimoto similarity measures — have been compared in previous work [14] and these same two measures are used in this study.

The most commonly employed similarity measure is a distance measure. Usually the Euclidean distance is used, but with binary data the same results are obtained from the Hamming distance, and the latter technique is more easily implemented on a computer. The Hamming distance between two spectra is obtained by applying an exclusive OR ("ORE") operation over all 139 dimensions. (See Table 1.) This procedure yields the number of mismatches between the two spectra (i.e., the Hamming distance). The unknown is predicted to belong to the same class as its nearest neighbor. The nearest neighbor of the unknown is the data set member for which the fewest mismatches (shortest Hamming distance) are obtained.

Two spectra that contain very few peaks will quite likely have fewer mismatches between them than two spectra that contain a larger number of peaks. The Tanimoto similarity measure [16] is calculated by normalizing the Hamming distance by the number of dimensions containing peaks in either or both spectra. The mathematics involved are shown in detail elsewhere [14]. The end results is that the Tanimoto similarity measure between spectra x and y, S_{xy}, is obtained by

$$S_{xy} = \sum_{j=1}^{139} (x_j \text{ AND } y_j) \Big/ \sum_{j=1}^{139} (x_j \text{ ORI } y_j) \tag{3}$$

The value of S_{xy} is minimized to determine the data set member that is most similar to the unknown.

ARTIFICIAL INTELLIGENCE APPROACH

The artificial intelligence approach attempts to parallel human reasoning as much as possible when interpreting a spectrum. Little work has been

TABLE 1

Boolean logic operators

ORE (exclusive OR)

	0	1
0	0	1
1	1	0

ORI (inclusive OR)

	0	1
0	0	1
1	1	1

AND

	0	1
0	0	0
1	0	1

reported on applying artificial intelligence techniques to infrared interpretation. Gray [17] used a purely empirical approach for interpreting spectra. The concepts used in the program discussed here are described below.

Basically, a set of rules for interpreting infrared spectra is determined. These rules may result from observation of a sufficient number of spectra or from reading textbooks and learning from the observations of others. For example, a simple set of rules for identifying carboxylic acids might be to look for a broad, medium to strong peak in the vicinity of 3000 cm^{-1}, a strong carbonyl peak near 1720 cm^{-1}, and a broad, medium intensity peak near 920 cm^{-1}. Similar rules are developed for all other functional groups of interest, and these rules are incorporated into the computer program. Next the program is tested on some infrared spectra. Whenever the program makes an erroneous interpretation, the rules must be altered to correct the mistake. As long as human interpretation of a spectrum is superior to the program, the program can be altered by using this new information so that it will then provide better interpretation. This ability to alter the program to correct mistakes is a major advantage of artificial intelligence programming.

One problem that must be faced first is how to encode the spectra. The major requirement for any encoding procedure is to allow the program logic to parallel human reasoning, which involves noting the positions of peaks as well as their intensities and shapes, but is not based on breaking the spectrum into intervals and noting the intensity for each interval. While the latter system of encoding spectra is probably best for retaining a maximum amount of information in a data set compilation, the method of encoding required by the infrared interpreter is similar to the one reported by Penski et al. [18]. For each peak, a code number indicating intensity and shape information is encoded along with the peak positions. The peak positions may be encoded in units of either cm^{-1} or μm. There are ten possible code numbers ranging from 0—9 with their meanings shown in Table 2. Admittedly, this method of encoding does not eliminate individual bias among chemists and will eventually require a rigid set of rules concerning what constitutes a strong peak vs. a medium peak or a broad peak vs. an average width peak. At this time, the distinction has been purposely left to discretion. Once the program has been used to test a large number of spectra,

TABLE 2

Intensity and shape codes
(Shoulder = 0)

		Intensity		
		Weak	Medium	Strong
Width	Sharp	1	4	7
	Average	2	5	8
	Broad	3	6	9

rigid guidelines will be developed. The program accounts for the possibility of some discrepancy in encoding peaks and this discrepancy should not harm the overall interpretation of spectra.

Theoretically, an artificial intelligence program which interprets infrared spectra could be written without reference to actual spectra. A training set or data set as described in the pattern recognition section is not needed. However, a collection of spectra is necessary in order to test the interpreter. As stated earlier, this testing may result in errors, which in turn will result in an improved program.

DATA SET

A total of 462 spectra were encoded by means of peak positions and the code numbers shown in Table 2. The spectra were encoded from a variety of sources, including textbooks by Silverstein and Bassler [19], Pasto and Johnson [20], and Nakanishi [21]. Four additional sources of spectra were the Sadtler collection [22], Umezawa's [23] index of antibiotics, an article by Hayden et al. [24], and a collection by Mecke and Langenbucher [25]. The 462 spectra were divided into three sets: 200 acids, 200 non-acids, and 62 which would be treated as unknowns for the pattern recognition work. The 200 acids were similar to the acid class used in previous pattern recognition work [10], thus their source was the ASTM file. Three of the ASTM spectra which were supposed to be of acids were actually esters, so that those three were replaced with acids. Since the ASTM collection contains only peak position information and no intensity or shape data, the original spectra had to be obtained (most were from Sadtler) and then encoded by hand. The reason for specifically choosing the same acids as were used in previous studies was that this choice allowed observations to be made concerning how well ASTM had encoded those spectra.

Observations on ASTM encoded spectra

One study [26] has used text-searching techniques to look for inconsistencies in the ASTM infrared file. While text searching was shown to be effective in finding inconsistencies in the molecular formula, compound name, and ASTM supplied classification codes, there was no way to look for errors in encoding of peak positions short of checking all entries by hand. While checking only 200 of the over 100,000 ASTM spectra by hand cannot result in any quantitative evaluation of an error rate in ASTM encoding, some observations can be made on the usefulness of the ASTM method of encoding for pattern recognition and artificial intelligence purposes.

Considering the enormity of the task of encoding over 100,000 infrared spectra, there are far fewer errors than one might expect, especially if the tolerance of ± 0.1 μm incorporated into most search systems is allowed. Errors do exist, however. For example, the spectrum of 6-methylpicolinic

acid (ASTM serial no. 15773CA) has all peak positions correctly encoded except one. The strong peak at 8.15 μm in that spectrum is omitted from the ASTM version. The spectrum of 1-naphthaleneacetic acid (10905CA) has more severe errors. Some of the intervals that contain peaks according to ASTM have no peaks in the original spectrum (e.g., 2.9—3.0 and 3.1—3.2 μm), while other intervals that do contain peaks in the original spectrum have no peaks indicated in the ASTM version (3.3—3.4, 8.2—8.3, and 10.7—10.8 μm). As indicated in the earlier study [26], some of the errors may be due not to faulty encoding, but to card-punching errors, as the original ASTM collection was on computer cards.

Perhaps more crucial than finding the few instances where specific errors exist in the ASTM file, is the consideration of whether the ASTM encoding policy of omitting shoulders and weak peaks [27] is wise. At least from the standpoint of artificial intelligence and pattern recognition programs, this policy will be shown to be unwise. If a chemist wanted to use only the peak positions present in the ASTM collection to interpret a spectrum, the task would be impossible in some instances even if intensity and shape information were available on those peaks that were indicated as present in the file. Often weak peaks are helpful in making an interpretation. For example, weak peaks between 2000—1660 cm^{-1} and sharp and often weak peaks around 1600 cm^{-1} and 1500 cm^{-1} are very helpful in identifying an aromatic compound. A specific example where weak peaks are vital for correct interpretation, is the spectrum of 4-methylpentanoic acid (2706CA). The spectrum has a relatively weak doublet at 7.23 μm and 7.36 μm. These peaks are not indicated in the ASTM file, yet they are essential if the *gem*-dimethyl functionality in the molecule is to be identified.

Identification of *gem*-dimethyl groups is also hindered because a resolution of 0.1 μm is insufficient at times. The spectrum of 2-ethyl-4-methylpentanoic acid (11286CA) has a doublet centered around 7.3 μm, yet the ASTM resolution is insufficient, so that only one peak can be indicated at 7.3 μm. Another hindrance to accurate interpretation of spectra if a resolution of 0.1 μm is used, is in determining the type of carbonyl functionality, e.g. distinction between a saturated aliphatic acid (5.81—5.86 μm) and an α,β-unsaturated acid (5.85—5.95 μm), is made considerably more difficult.

Finally, the policy of ignoring shoulders also makes the task of interpreting an infrared spectrum more difficult. Spectra of carboxylic acids usually contain a shoulder near 3.8 μm. Identification of many other functionalities would be simplified if shoulder presence were indicated.

To summarize the observations on the ASTM file, inconsistencies do exist, as will always be the case when human judgement is required. Two specific comparisons illustrate this point. The ASTM entry for 3,3-dimethylheptanoic acid (4194CA) contains a peak in the 3.7—3.8 μm region. The vast majority of the other 200 acid spectra examined have a shoulder that is just as evident in this region, yet their ASTM entries indicate no peak in that interval. As a general rule, the ASTM entries for spectra run in mineral oil

do not indicate peaks caused by the oil (e.g., between 3.4 and 3.5 μm). However, the spectrum of myristic acid (2807CA) does have the mineral peak encoded.

Numerous papers have described successful search systems employing the ASTM collection. One reason for their success is that the ASTM file contains many duplicate entries. For example, there are at least sixteen cyclohexanone entries. Thus, if the unknown spectrum were of cyclohexanone, the chances of making a correct match are good, even if one of the cyclohexanone entries contains errors.

Some observations about the ASTM method of encoding infrared spectra as they pertain to pattern recognition will be presented in the next section. Even if the ASTM collection were to include intensity and shape information on the peaks, the collection would be of little value as a test set for an artificial intelligence program. Whether human intelligence or an artificial intelligence program is used, information about weak peaks and shoulders is often required, and better than 0.1-μm resolution is needed for interpretation.

PATTERN RECOGNITION RESULTS

The 462 data set members were divided into two groups, a 400-member training set (200 acids and 200 non-acids) and a group of 62 unknowns. Experiments were performed to attempt to answer two questions. Does intensity and shape information improve classification ability? Do shoulders improve classification ability? Each of the 462 spectra was encoded as required by the artificial intelligence program, i.e., peak positions and code numbers. To represent each spectrum as a 139-dimensional vector for the pattern recognition work, the peak positions were rounded off to the nearest 0.1 μm.

The recognition results from the training set are shown in Table 3. Several trends are evident from these results. Similarly to previous work [10], the average classification ability for distance from the mean is greater than for Tanimoto similarity which is greater than for nearest neighbor. Also similarly

TABLE 3

Percentage recognition of training set

Type of spectral data	Distance from mean			Nearest neighbor			Tanimoto similarity		
	Acid	Non-acid	Av.	Acid	Non-acid	Av.	Acid	Non-acid	Av.
Binary (inc. shoulders)	89.0	95.0	92.0	72.0	94.5	83.2	90.5	84.5	87.5
Binary (no shoulders)	81.5	86.5	84.0	68.0	83.5	75.7	81.0	79.5	80.2
Intensity and shape	79.5	99.0	89.2	70.5	97.5	84.0	—	—	—

to previous findings [11—13], binary data and data that included intensity information are classified about equally well (binary 2.8% better than intensity for distance from the mean; intensity 0.8% better than binary for nearest neighbor). Finally, classification of binary data that retained shoulders is significantly better than for binary data without shoulders (the method used by ASTM). For the three techniques investigated, the data with shoulders are classified correctly between 7.3 and 8.0% more frequently than the data without shoulders. In fact, 83.5% of the 400 training set members are recognized correctly with the distance-from-the-mean technique when only the peak positions of the shoulders are used and all other peaks are ignored. The 83.5% figure compares favorably with the figure of 84.0% for binary data without shoulders. Thus, it appears that shoulders do aid classification.

The prediction results for the 62 unknowns are shown in Table 4. While the average results in all cases are poorer than for recognition of the training set, the same trends as described above are observed. No definite superiority is demonstrated for binary data vs. intensity and shape data. However, once again the data with shoulders give better results than data without shoulders.

ARTIFICIAL INTELLIGENCE RESULTS

It is most important for the infrared interpreter to make decisions about a large number of functionalities. The program investigates 169 different classes, divided into major classes (aromatic, acid, ester, amine, etc.) and subclasses (monosubstituted benzene, acetate, α,β-unsaturated acid, etc.). Reference 1 includes a complete listing of the classes and results from a study on 243 different spectra. Obviously, no program will give a correct yes/no answer for 169 classes all of the time. However, the entire structure elucidation scheme is not based exclusively on infrared data. Obviously, supplementary information will be used to deduce the correct structure. The infrared interpreter is designed to help in full structural elucidation. The program provides one of five possible confidence levels for each of the 169 classes: 0, definitely absent; 1, low probability; 2, medium probability;

TABLE 4

Percentage prediction for unknowns

Type of spectral data	Distance from mean			Nearest neighbor			Tanimoto similarity		
	Acid	Non-acid	Av.	Acid	Non-acid	Av.	Acid	Non-acid	Av.
Binary (inc. shoulders)	82.4	92.9	87.6	38.2	100.	69.1	79.6	80.4	80.0
Binary (no shoulders)	61.8	75.0	68.4	47.1	80.4	63.7	64.7	75.0	69.9
Intensity and shape	61.8	100.	80.9	47.1	100.	73.5	—	—	—

3, high probability; 4, definitely present. The worst situations would be for the program to return a confidence value of 4 (definitely present) if the functionality were absent or a 0 for a functionality that is present. The method of selecting confidence levels differs from class to class, but in accordance with artificial intelligence programming, the selection is designed to simulate human reasoning.

For the purpose of comparison with the pattern recognition results, only the carboxylic acid class and its four subclasses are considered below. The acid subclasses are α,β-unsaturated, no α,β-unsaturation (saturated), α-electronegative group attached, and pyridine-like acid (showing broad absorption near 2450 and 1900 cm^{-1} and no broad absorption near 3000 and 920 cm^{-1}). The 400 spectra used as the training set for the pattern recognition work were tested with the original program [1]. The original program had been tested on only 17 acids. All 200 non-acids resulted in confidence of zero for the acid class. Twenty of the 200 acid spectra resulted in confidences of either 0 or 1 for the acid class (fifteen 0's and five 1's) with the original program. Thus, 7.5—10.0% of the acid class were incorrectly interpreted, depending on whether or not a confidence level of 1 was considered to be an error. The average figure of 95% for the 400 acids was only slightly superior to the distance from the mean results. However, a major difference between artificial intelligence and pattern recognition programming is that with the former, the program can be changed to improve performance if a chemist can determine which interpretation rules are faulty. Most of the errors resulted from inadequate rules for interpreting pyridine-like acids. The infrared interpreter was modified based on the errors observed. The present version still correctly identified the 200 non-acids and yet found only two of the acids to have a confidence of 0 and two acids with a confidence of 1.

The four problem compounds are all pyridine-like acids which appear to be most stable as zwitterions. Thus, the faulty rules are among the carboxylate anion class rules and not the carboxylic acid class rules. Further investigation of zwitterions is necessary to correct the anion class rules.

Tests on the 62 spectra used as unknowns in the pattern recognition studies gave the following results. All 28 non-acid spectra were correctly interpreted as being non-acids, whereas 6 of the 34 (17.6%) acids were incorrectly intepreted as being non-acids. Once again, an experienced chemist would have performed better. The errors for the most part were due to the carbonyl absorption appearing at a slightly higher frequency than the rules allowed. The program had not previously been tested on compounds like trifluoroacetic acid, so that the rules were too restrictive. After another modification of the program, all 62 unknowns are now correctly interpreted.

CONCLUSIONS

The ASTM file of infrared spectra has some encoding inconsistencies. While numerous studies have found the ASTM file to be adequate for search systems, especially since the file contains much duplication, the ASTM file does have some drawbacks for pattern recognition work and it is of minimal value for artificial intelligence programming. That the file contains binary data does not appear to be detrimental to pattern recognition work. The results found here agree with previous studies that inclusion of intensity data does not alter classification ability significantly. However, significant improvements in both recognition and prediction results are obtained if data that retain the positions of shoulders are used rather than disregarded, as in ASTM policy.

The percentage of spectra correctly interpreted by the artificial intelligence program before modifications, was similar to the figure correctly recognized by the distance-from-the-mean technique. A similar situation existed for the 62 unknown spectra. Both techniques can be refined. For an artificial intelligence program, all that is required is to alter the rules whenever human interpretation is superior to the program. Classification ability by pattern recognition ought to be improved by using substantially larger training sets. The assumption is that the larger the training set, the more representative it is of the real world. Of course, this method of refinement is not easily achieved, as training is a slow process and sufficient spectra properly encoded may not be available.

To be valuable for structural elucidation, it is not sufficient for a computerized interpretation procedure merely to distinguish between acids and non-acids; the kind of acid must be known. While it is theoretically possible for pattern recognition procedures to test 169 classes of compounds, the size of the training set and the investment of computer time to complete training would be enormous.

Thus, for the specific task of designing a program to be used in a structural elucidation scheme, the artificial intelligence approach is the method of choice. The artificial intelligence program can work with a larger number of functionalities more easily than pattern recognition; it can be modified more easily; it can be a valuable tool in helping to plan new experiments. The present infrared interpreter is more successful in identifying acids than the original version; after extensive testing of actual infrared spectra has been done, the interpreter will be even more valuable for solving structural elucidation problems.

The authors acknowledge the support of this project by the National Institutes of Health (GM 21703).

REFERENCES

1 H. B. Woodruff and M. E. Munk, J. Org. Chem., 42 (1977) 000.
2 H. Abe and S. Sasaki, Sci. Rep. Tohoku Imp. Univ., Ser. 1, 55 (1972) 63.
3 B. D. Cox, Ph.D. Thesis, Department of Chemistry, Arizona State University, 1973.
4 L. A. Gribov, V. A. Dementyev, M. E. Elyashberg, and E. Z. Yakupov, J. Mol. Struct., 22 (1974) 161.
5 R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, J. Am. Chem. Soc., 97 (1975) 5755.
6 C. A. Shelley, H. B. Woodruff, and M. E. Munk, Abstracts, Division of Chemical Information, 173rd National American Chemical Society Meeting, New Orleans, March 23, 1977.
7 B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, Anal. Chem., 41 (1969) 1945.
8 D. R. Preuss and P. C. Jurs, Anal. Chem., 46 (1974) 520.
9 R. W. Liddel III and P. C. Jurs, Anal. Chem., 46 (1974) 2126.
10 H. B. Woodruff, G. L. Ritter, S. R. Lowry, and T. L. Isenhour, Appl. Spectrosc., 30 (1976) 213.
11 S. L. Grotch, Anal. Chem., 47 (1975) 1285.
12 J. B. Justice and T. L. Isenhour, Anal. Chem., 46 (1974) 223.
13 H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, Appl. Spectrosc., 29 (1975) 226.
14 H. B. Woodruff, S. R. Lowry, G. L. Ritter, and T. L. Isenhour, Anal. Chem., 47 (1975) 2027.
15 R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience, New York, N.Y., 1973.
16 D. J. Rogers and T. T. Tanimoto, Science, 132 (1960) 1115.
17 N. A. B. Gray, Anal. Chem., 47 (1975) 2426.
18 E. C. Penski, D. A. Padowski, and J. B. Bouck, Anal. Chem., 46 (1974) 955.
19 R. M. Silverstein and G. C. Bassler, Spectrometric Identification of Organic Compounds, 2nd edn., J. Wiley, New York, 1967.
20 D. J. Pasto and C. R. Johnson, Organic Structure Determination, Prentice-Hall, Engelwood Cliffs, N.J., 1969.
21 K. Nakanishi, Infrared Absorption Spectroscopy, Holden-Day, San Francisco, Calif., 1962.
22 Sadtler Standard Infrared Spectra, Sadtler Research Laboratories, Philadelphia, Pa.
23 H. Umezawa, Index of Antibiotics from Actinomycetes, University Park Press, State College, Pa., 1967.
24 A. L. Hayden, O. R. Sammul, G. B. Selzer, and J. Carol, J. Assoc. Off. Agric. Chem., 45 (1962) 797.
25 R. Mecke and F. Langenbucher, Infrared Spectra of Selected Chemical Compounds, Heydon, London, 1970.
26 J. A. de Haseth, H. B. Woodruff, and T. L. Isenhour, Appl. Spectrosc., 31 (1977) 18.
27 Codes and Instructions for Wyandotte-ASTM Punched Cards, Indexing Spectral Absorption Data, American Society for Testing and Materials, Philadelphia, Pa., 1964.

Analytica Chimica Acta, 95 (1977) 25—32
Computer Techniques and Optimization

COMPUTER-AIDED INTERPRETATION OF STEROID MASS SPECTRA BY PATTERN RECOGNITION METHODS§

Part 2. Influence of Mass Spectral Preprocessing on Classification by Distance Measurement to Centres of Gravity.

H. ROTTER and K. VARMUZA*

Institut für Allgemeine Chemie, Technische Universität, Lehargasse 4, A-1060 Vienna (Austria)

(Received 31st January 1977)

SUMMARY

Seventeen preprocessing methods have been applied to 524 low-resolution mass spectra of steroids before computing classifiers, which can recognize substructures in a steroid molecule. Best classification results have been obtained by normalization of peak height to local ion current (predictive abilities 85%) and with "significant" spectra that contain only the "most important" peaks (predictive abilities 84%).

In recent years pattern recognition methods have been used for the interpretation of low-resolution mass spectra [2]. In most pattern recognition methods, classifiers are derived from a random sample of spectra which stem from substances of known structure. A classifier can be understood as an algorithm that assigns a spectrum to one of several possible classes. Each class corresponds to a certain chemical structure. Binary classifiers distinguish between 2 classes. They have been applied successfully in interpreting mass spectra of low-molecular-weight substances [2, 3] and steroids [1, 4].

The first step during the development of a classifier is spectral preprocessing. This is done by extracting features from mass spectral data; the features of one spectrum are then combined to a pattern vector. A feature is a function of the recorded mass spectral data (mass numbers and peak heights). It is desirable to use only a few features with as large an amount of information as possible. A small number of features results in short computing time and relatively little computer storage requirement. For low- resolution mass spectra, a simple way to deduce features is the following widely used procedure: peak heights at mass number m are equivalent to components (features) with number m of the pattern vector. Instead of peak heights, h_m, simple functions of h_m can be used (e.g. logarithms). To reduce the number of dimensions in pattern vectors, the most important features may be selected, but feature selection is not dealt with here.

---

§ For Part 1, see [1].

Spectral preprocessing has a great influence on classification results, but in general the features which give best results cannot be specified a priori. Rather, different preprocessing methods must be tested with a certain classification procedure. The best spectral preprocessing is that which gives the best results for classification. This optimal spectral preprocessing will be valid only for a certain classification procedure and only for that set of mass spectra from which the random sample stems. Nevertheless, qualitative conclusions may be drawn also for other methods. This paper reports on preprocessing of low-resolution mass spectra from steroids, and distance measurement to centres of gravity was used throughout this work as the classification method.

DATA AND PROGRAMMES

The spectral file consisted of 524 low-resolution mass spectra from 524 different steroids [5]. Elemental composition of the steroids was in the range $C_{18-29}H_{18-50}N_{0-1}O_{0-8}F_{0-3}Si_{0-2}$, molecular weights between 256 and 463, and mass ranged from 39 to 463. The spectral file was divided into 2 classes in 17 different ways. Class 1 always contained all steroids that have a certain chemical substructure, and class 2 contained all other spectra (Table 1). Programmes were written in FORTRAN and were run on a CDC Cyber 74

TABLE 1

Substructures of steroids
(N_1 is the number of steroids in the file that contain a certain substructure; $p(1)$ is the corresponding probability.)

No.	Substructure	N_1	$p(1)$
1	Double bond C=C	238	0.45
2	Double bond C-4=C-5	129	0.25
3	Hydroxysteroid	380	0.73
4	Ketosteroid	382	0.73
5	Oestrane- or androstane-type	293	0.56
6	3-Hydroxysteroid	204	0.39
7	3-Ketosteroid	237	0.45
8	Oxygen function at C-3	479	0.91
9	Oxygen function at C-11	115	0.22
10	Oxygen function at C-17	330	0.63
11	5α-Steroid	197	0.38
12	5β-Steroid	129	0.25
13	OH in side chain at C-17	66	0.13
14	CO in side chain at C-17	120	0.23
15	20-Ketopregnane	119	0.23
16	20-Hydroxypregnane	30	0.06
17	Carboxyl group	56	0.11

(memory, 98K words, 60 bits) at the Computer Centre, University of Technology, Vienna.

CLASSIFICATION BY DISTANCE MEASUREMENT TO CENTRES OF GRAVITY

The pattern vector of a spectrum represents a point in an n-dimensional spectral space. It is assumed that spectra of substances with similar chemical structure have similar pattern vectors and therefore form clusters in the spectral space. In a simple case each of the classes 1 and 2 corresponds to one cluster. Each cluster may be characterized by one prototype point. The centre of gravity (which is equivalent to an averaged pattern vector of all spectral points which belong to this class) was chosen as the prototype point [6, 7]. To classify an unknown spectrum, the distances to the prototype points are calculated. The spectrum is then coordinated to the class with the smallest distance. In this paper, the Euclidean distance is used. For binary classifiers, one always has 2 classes with 2 corresponding centres of gravity which define a linear classifier. Distance measurement to prototype points is a standard method in pattern recognition. It is a simple and clear method and is easily applicable even if pattern vectors contain several hundred dimensions. These are the reasons why this method has been used here for an extensive comparison of spectral preprocessing, although there are other classification methods which give better results.

For each of the 17 substructures listed in Table 1, a classifier was computed from all 524 mass spectra. To test a classifier the same 524 spectra were classified, and criteria were computed that characterize a binary classifier objectively. As mentioned before, predictive abilities P_1 and P_2 together [8] or maximum information [9], I_{max}, are suitable for this purpose. P_1 and P_2 are the percentages of correctly classified spectra of classes 1 and 2, respectively; I_{max} is the maximum information in bits that is obtained if the classifier is used with a sample of spectra that has equal numbers of members in each class. These criteria make it possible to select the best preprocessing method for each substructure. To select a preprocessing method suitable for all structures, averaged values ($\bar{P}_1, \bar{P}_2, \bar{I}_{max}$) from P_1, P_2 and I_{max} were calculated. It is reasonable to use the same spectra for classifier development and for classifier testing because only the statistical properties (i.e. mean values) of the spectral file are important. Computations showed that the position of a centre of gravity does not change significantly if only one half of the spectral file is considered.

SPECTRAL PREPROCESSING METHODS

Intensities normalized to the base peak were used directly as features for pattern vectors (normal spectra). For logarithmic spectra, features x_m were calculated from peak heights h_m:

$$x_m = (100 \log h_m + 100)/3 \text{ if } h_m > 0.1\% \text{ base peak}$$
$$x_m = 0 \qquad \text{if } h_m \leqslant 0.1\% \text{ base peak} \qquad (1)$$

This transformation again yields features in the intensity range 0—100. In binary encoded spectra [10], features x_m are 1 if peak heights h_m are greater than a threshold, otherwise x_m are 0. Thresholds of 0.05 and 5% base peak were used.

Another preprocessing is normalization to total ion current (which is equivalent to the sum of the peak heights). Normalization to local ion current [11, 12] was done by dividing each peak height by the sum of peak heights in a window of several mass units. Features x_m were calculated from peak heights h_m with a window width of $2\Delta m + 1$:

$$x_m = h_m / \sum_{j=m-\Delta m}^{m+\Delta m} h_j \qquad (2)$$

Normalization to local ion current enlarges isolated peaks or isolated peak groups. Figure 1 shows the normal mass spectrum (a) of 3β-hydroxy-5β-androstan-17-one and the mass spectrum (b) normalized to local ion current ($\Delta m = 3$).

Selection of the "most important" peaks was done by an heuristic method proposed by Nägeli and Clerc [12]. Figure 2 shows the procedure which was used to eliminate "unimportant" peaks in order to generate significant spectra. In binary significant spectra, peak heights of all significant peaks were set to value 1. Significant spectra with normalization to local ion current were also used. These spectra contained only significant peaks according to Fig. 2, but peak heights were taken from full spectra which had been normalized to local ion current. An example of such a spectrum is shown in Fig. 1c.

From information theory, it is known that a signal contains most information if all possible values of the signal have equal probability. To generate pattern vectors with equally distributed features from mass spectra, intensity levels must be created for each mass number. Each level should have the same probability in the spectral file. If there are L intensity levels and N mass numbers, a set of $(L-1)N$ threshold values must be stored for spectral preprocessing. Equally distributed features are necessary for the classification methods proposed by Franzen [13] and Hillig [14]. Pattern vectors with 8 intensity levels were used in this paper.

Data reductions which are widely used in library search systems [15] were also used for preprocessing. One method is "k largest peaks in spectrum" ($k = 5, 10, 20, 50$); the other method is "k largest peaks in mass intervals of length l" ($k = 2$, $l = 14$ and $k = 1$, $l = 7$).

Another preprocessing method tries to give equal importance to all features. If one feature has values in the range 0—1 and another feature in the range 0—100, the second feature is more important in almost all classification methods. This effect may be compensated by "autoscaling" [2, 16]. If s_m is the standard deviation of feature h_m (computed over all spectra), an autoscaled feature x_m is obtained from

$$x_m = h_m / s_m \qquad (3)$$

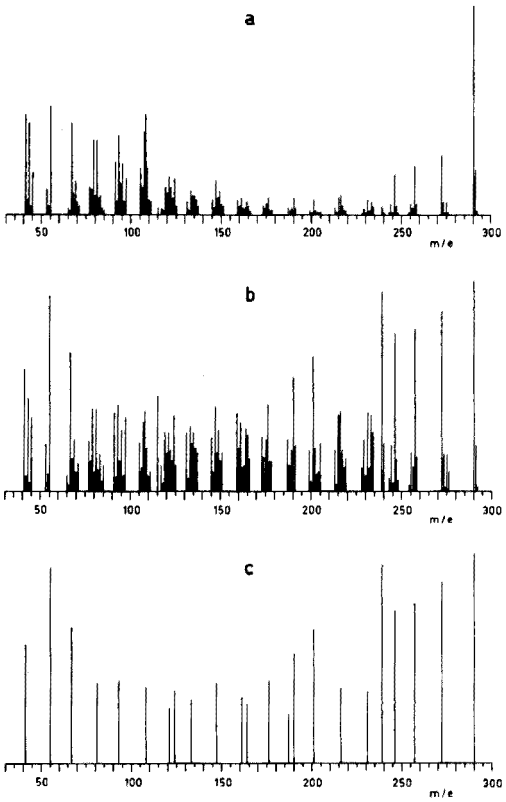


Fig. 1. Mass spectra from 3β-hydroxy-5β-androstan-17-one after different preprocessing. (a) Normal spectrum. (b) Normalization to local ion current (Δm = 3). (c) Significant spectrum normalized to local ion current (Δm = 3). Peak heights in arbitrary linear units.

RESULTS AND DISCUSSION

To characterize the quality of a preprocessing method, the classification results for 17 chemical structures were averaged. Mean values $\overline{P}_1$, $\overline{P}_2$, $\overline{I}_{max}$ for all preprocessing methods are summarized in Table 2; there are only small differences between the results derived from mean values. Logarithmic spectra, binary spectra, and spectra with equally distributed intensities do not improve the classification results in comparison with normal spectra. Significantly better are the spectra normalized to total ion current.

The best classification results of all preprocessing methods are obtained if spectra are normalized to local ion current. Figure 3 shows the dependence of $\overline{I}_{max}$ on window width. Best results are obtained for Δm = 3 (i.e. the window width is 7 mass units). Table 3 contains detailed results for all 17 chemical structures for this preprocessing method. Predictive abilities for classes 1 and 2 lie between 67% and 97% with an averaged value of 85%.

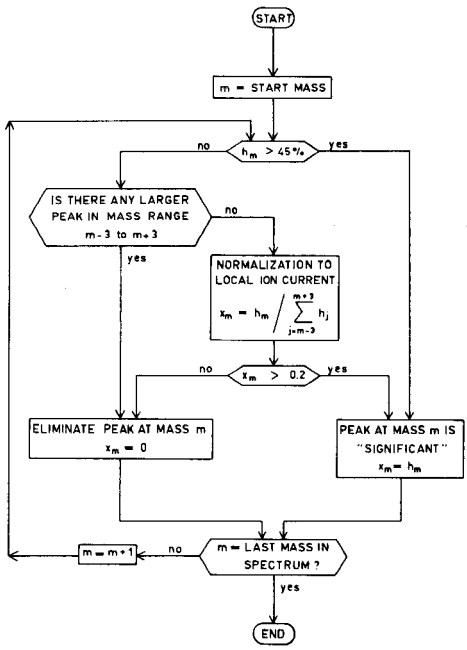


Fig. 2. Program scheme for generating significant spectra [12] (with peak heights x_m) from normal spectra (peak heights h_m in % base peak).

Significant spectra give better results than normal spectra; an improvement is possible, if significant spectra contain only binary encoded peak heights or if peak heights are normalized to local ion current. The slight improvement of classification after autoscaling does not justify the computational effort necessary.

Surprisingly good results are obtained if data reduction is applied to some of the largest peaks. Classification is similar to normal (full) spectra, but is less satisfactory than with other preprocessing methods. This agrees with the results of library search experiments done with the same spectral file [17].

It can be concluded that classification by distance measurement is predominantly controlled by the largest peaks in a spectrum. Normalization to local ion current gives the best results for classification by distance measurement to centres of gravity, but has the disadvantage that data reduction is not combined with preprocessing. The use of significant spectra with normalization to local ion current reduces the number of peaks per spectrum significantly while classification results stay very satisfactory. The good results obtained with significant spectra show that chemical experience gives valuable hints for feature selection, which is a central problem in automated spectral interpretation by pattern recognition methods.

TABLE 2

Classification results
($\overline{P_1}$, $\overline{P_2}$ and $\bar{I}_{max}$ are averaged values for 17 chemical structures.)

Preprocessing	$\overline{P_1}$	$\overline{P_2}$	$\bar{I}_{max}$ (bit)
Normal spectra	0.77	0.80	0.273
Logarithmic spectra	0.77	0.78	0.249
Binary spectra (threshold 0.05%B)	0.76	0.77	0.234
Binary spectra (threshold 5%B)	0.76	0.77	0.223
Norm. to total ion current	0.82	0.80	0.316
Norm. to local ion current (Δm = 3)	0.86	0.84	0.417
Significant spectra	0.79	0.80	0.290
Binary significant spectra	0.84	0.81	0.350
Sign. spectra norm. to local ion current (Δm = 3)	0.85	0.83	0.394
Spectra with equally distr. features (8 levels)	0.75	0.75	0.205
Autoscaled spectra	0.78	0.83	0.320
5 largest peaks	0.77	0.77	0.245
10 largest peaks	0.79	0.81	0.296
20 largest peaks	0.78	0.80	0.280
50 largest peaks	0.78	0.80	0.277
2 largest peaks in intervals of 14	0.79	0.81	0.295
1 largest peak in intervals of 7	0.79	0.79	0.283

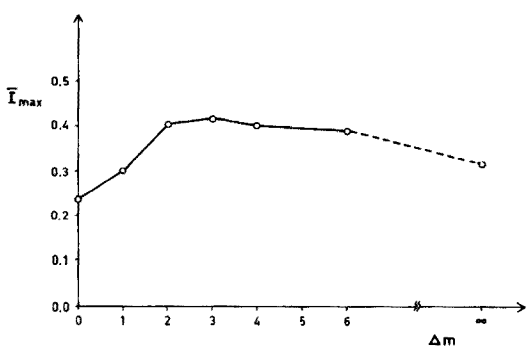


Fig. 3. Maximum information $\bar{I}_{max}$ (in bit) for normalization to local ion current versus window width, $2\Delta m + 1$. $\Delta m = 0$ corresponds to a binary encoded spectrum with a threshold of 0.05% base peak; $\Delta m = \infty$ corresponds to normalization to total ion current.

We thank Prof. Dr. G. Spiteller for the use of his library of steroid mass spectra, Prof. Dr. A. Maschka for his support of this work, and Mr. H. Urban for technical assistance.

TABLE 3

Classification results for normalization to local ion current (window width 7 mass units)

No.	Substructure	P_1	P_2	I_{max} (bit)
1	Double bond C=C	0.83	0.92	0.467
2	Double bond C-4=C-5	0.83	0.88	0.401
3	Hydroxysteroid	0.81	0.84	0.326
4	Ketosteroid	0.83	0.83	0.346
5	Oestrane- or androstane-type	0.90	0.92	0.569
6	3-Hydroxysteroid	0.76	0.81	0.244
7	3-Ketosteroid	0.79	0.82	0.287
8	Oxygen function at C-3	0.86	0.76	0.299
9	Oxygen function at C-11	0.82	0.82	0.316
10	Oxygen function at C-17	0.86	0.86	0.404
11	5α-Steroid	0.80	0.73	0.215
12	5β-Steroid	0.80	0.67	0.165
13	OH in side-chain at C-17	0.96	0.87	0.588
14	CO in side-chain at C-17	0.94	0.89	0.590
15	20-Ketopregnane	0.94	0.89	0.585
16	20-Hydroxypregnane	0.97	0.95	0.750
17	Carboxyl group	0.91	0.89	0.535

REFERENCES

1 Part 1: K. Varmuza, H. Rotter and P. Krenmayr, Chromatographia, 7 (1974) 522.
2 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
3 K. Varmuza, Mh. Chem., 107 (1976) 43.
4 K. Varmuza, Fresenius' Z. Anal. Chem., (1977), in press.
5 G. Spiteller, Steroid Mass Spectra Library, Göttingen.
6 G. M. Meyer-Brötz and J. Schürmann, Methoden der automatischen Zeichenerkennung, Oldenburg, München-Wien, 1970.
7 W. S. Meisel, Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York and London, 1972.
8 H. Rotter and K. Varmuza, Org. Mass Spectrom., 10 (1975) 874.
9 K. Varmuza and H. Rotter, Mh. Chem., 107 (1976) 547.
10 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
11 J. T. Clerc, private communication.
12 P. R. Nägeli, Dissertation, ETH Zürich, 1975.
13 J. Franzen, Chromatographia, 7 (1974) 518.
14 H. Hillig, Dissertation, Univ. Dortmund, 1975.
15 R. G. Ridley, in G. R. Waller (Ed.), Biochemical Applications of Mass Spectrometry, Wiley, New York, 1972.
16 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.
17 K. Varmuza, Fresenius' Z. Anal. Chem., 282 (1976) 129.

Analytica Chimica Acta, 95 (1977) 33—40
Computer Techniques and Optimization

A MINICOMPUTER PROGRAM BASED ON ADDITIVITY RULES FOR THE ESTIMATION OF ^{13}C-NMR CHEMICAL SHIFTS

J. T. CLERC* and H. SOMMERAUER

Laboratorium für organische Chemie, Eidgenössische Technische Hochschule, Zürich, (Switzerland)

(Received 1st April 1977)

SUMMARY

A minicomputer program for the estimation of ^{13}C-n.m.r. chemical shifts based on simple additivity rules is described.

When ^{13}C-n.m.r. data are used for the structural elucidation of organic compounds, the analyst is often confronted with the problem of predicting chemical shift values for compounds of various structures. If spectra of suitable reference compounds are available, they may be used directly [1—3]; otherwise the data may be estimated using additivity rules. The use of additivity rules is well established in ^{1}H-n.m.r. [3, 4]. Similar rules have been developed for the estimation of ^{13}C-n.m.r. chemical shifts [3, 5, 6]. Generally, for the prediction of ^{1}H-n.m.r. shifts, only contributions from α-neighbours are taken into account. ^{13}C-n.m.r. shifts may, however, be significantly influenced even by substituents in δ-position. Thus a comparatively large number of increments must be summed, so that manual estimation becomes tedious and error-prone even with rather simple compounds. Since all modern ^{13}C-n.m.r. spectrometers are equipped with minicomputers, it seems feasible to develop appropriate computer programs to avoid this task. This paper describes a computer program which will predict from a given chemical structure, the ^{13}C-n.m.r. chemical shift values by means of additivity rules. In its present version the program is limited to singly bonded carbons in acyclic compounds that contain any combination of noncyclic functional groups listed earlier [7]. The implementation on a BRUKER BNC 28 minicomputer uses 8K core storage. The processing speed is input-limited. A high-level language version (FORTRAN IV) is under development.

Overview

Under program control the following operations are performed sequentially. First, the user inputs the structure to be operated on. He enters the skeleton by specifying linear chains of arbitrary length; such chains may be grafted at arbitrary positions on chains already input, thus allowing for easy generation

of branched structures. The skeleton is implicitly assumed to represent a saturated hydrocarbon. In the following step, nodes in the skeleton may be changed to represent heteroatoms instead of carbons, and edges may be defined to correspond to multiple bonds rather than to single bonds. Thus arbitrary complex structures of acyclic compounds may easily be input with just three different command types. Internally, the structure is represented as a redundant connection table (cf. Table 2).

In the next phase, the program identifies the functional groups present. It operates on the connection table, assigning to each node a special code which gives the type of the corresponding atom, the number and type of its neighbours and the type of bonds emanating from it. Then the perceived structure is output in linear form, together with the numbering scheme used. This serves as a check for the structure input. Then for every singly bonded carbon atom the increments for its neighbours are retrieved from a table and summed. Finally, the estimated chemical shift values are output in tabular form. An example is given in Table 1. A more detailed description of selected program modules follows.

The accuracy of the estimated chemical shift values is obviously determined by the limitations of the linear model used. However, the program allows for greatly enhanced speed and user comfort. Furthermore, in combination with large computer-readable collections of ^{13}C-n.m.r. reference data, it makes feasible the use of sophisticated optimization procedures for the determination

TABLE 1

Sample output from ^{13}C-n.m.r. estimation program
(User interactions underlined)

```
C13NMR ESTIMATION

#K 1 4
#K 5 5 2
#H 0 4
#

STRUCTURE

1 2 [5] 3 4

CH3-CH[—CH3]—CH2—OH

CALC SPECTRUM

ATOM NR     SHIFT

     1       18.3
     2       31.4
     3       72.1
     5       18.3
```

of substituent increments. Selected entries from such a collection may also be used to provide base values for those compound classes for which the simple linear model does not give acceptable results. Preliminary studies have shown that the linear approximation may be extended to substituted cyclic compounds if the shift values of the parent ring compound are used as base values. A limited set of ring skeletons with their associated shift values will later be incorporated into the data tables. These will provide base values for the estimation of chemical shift values in cyclic compounds. Furthermore, a standard drawing of the ring skeleton for output on a character-oriented display system will be stored with each base ring system. These templates serve as a starting point for assembling a topological representation of the structure to be processed for output on a character-oriented terminal.

PROGRAM DESCRIPTION

Structure input

For structure input only three different commands are needed. The first command, K, is used to enter unbranched chains. The command K carries three arguments, namely the number of the first and the last atom of the chain and the position where the new chain is connected to the skeleton already input. For the first chain, the third argument is not specified. The command H is then used to replace carbon atoms by heteroatoms. As arguments, the type of the heteroatom and its position have to be specified. Finally, multiple bonds may be entered with the command B, followed by the multiplicity of the bond and the numbers of the atoms it connects. The conversion of a sequence of such commands into a redundant connection table is a straightforward operation. As only three different commands are used, this method for inputting structures is very simple to learn and to apply. Input of cyclic structures is easily accomplished by adding one more command to denote ring closures. To enter highly branched structures many short chains have to be combined, which results in a rather voluminous input sequence. This slight disadvantage is, however, more than balanced by the very simple and obvious command structure. An example of the input of a relatively simple compound is given in Table 2.

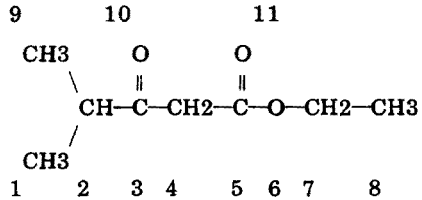
Perception of functional groups

After building the connection table from the user's input commands, the program continues by identifying the functional groups present. Initially each node in the structure is identified with the atom type it represents. From these codes new codes are generated, describing the type of the central atom, the types of its immediate neighbours and their type of bonding. This is done as follows. Each atom type is associated with a number which encodes its type and the multiplicity of its bond to the central atom. These numbers are selected to give a unique sum for each distinct combination of α-neighbours. For ease of processing, this code sum is split into two code words. The first

TABLE 2

Structure input and internal representation

Sample structure:



Input commands:
C13NMR ESTIMATION

```
#K 1 8
#K 9 9 2
#K 10 10 3
#B 2 3 10
#H 0 10
#K 11 11 5
#B 2 5 11
#H 0 11
#H 0 6
#
```

Internal representation (connection table)

Atom No.	Type	1. Neighbour		2. Neighbour		3. Neighbour	
		No.	Bond	No.	Bond	No.	Bond
1	C	2	1				
2	C	1	1	3	1	9	1
3	C	2	1	4	1	10	2
4	C	3	1	5	1		
5	C	4	1	6	1	11	2
6	O	5	1	7	1		
7	C	6	1	8	1		
8	C	7	1				
9	C	2	1				
10	O	3	2				
11	O	5	2				

word holds the sum for all singly bonded neighbours, whereas the codes for all multiply bonded neighbours are accumulated in the second word. In addition, the type of the central atom is also encoded in this second part. A partial list of these code numbers is reproduced in Table 3. The code sums generated for a simple compound are given later.

In a first step, these code sums are checked against a list of those codes that unambiguously identify a functional group. A part of this list is reproduced in Table 4. If a code is found in that list, the corresponding node is marked by replacing the atom type in the connection table with a pointer to the shift increments corresponding to the respective functional group. If no hit results from the scanning of this list, a second list is inspected, which gives those code sums that correspond to those functional groups that may be unambiguously identified if the β-neighbours are taken into consideration too (cf. Table 5). If a hit is scored here, the codes of the α-neighbours are inspected and again the respective node is marked (in some instances, also its neighbours). If no corresponding entry is found, the program proceeds without any marking of nodes. When all nodes in the structure are processed as described, the program checks to see if all nodes have been marked. If this is the case, all functionalities have been perceived. Otherwise the structure contains some functional group for which no increments are known. The program then generates a message and aborts. The perception of functional groups for a simple compound is summarized in Table 6.

Structure output

To guard against input errors, the structure perceived is again output for checking purposes. As errors made in transforming a graphical representation to a string of input commands tend to be reproduced, the output must be different from the input. Furthermore, it should be easily interpretable by the user. Since neither a record of the input commands nor the connection

TABLE 3

Code increments for central atom type and α-neighbours

First code word		Second code word			
Neighbour atom type	Code increment (octal)	Neighbour atom type	Code increment (octal)	Central atom type	Code increment (octal)
—C	1	=N	1	C	100000
—O	5	≡N	3	O	200000
—S	25	=S	6	S	300000
—N	105	=O	20	N	400000
—P	417	=P	47	P	500000
—F	1251	=C	70	F	600000
—Cl	5244	≡C	230	Cl	700000
—Br	25220			Br	1000000
—I	125100			I	1100000

TABLE 4

Single codes identifying functional groups (octal values)[a]

Part 1	Part 2	Functionality
Any value	100000	Aliphatic C
1	100020	Aldehyde C
2	100020	Ketone C
1	100003	Nitrile C
0	100070	H2C=
1	100070	—HC=
0	100230	HC≡
1	100230	—C≡
0	200070	Carbonyl O
0	400230	Nitrile N
1	600000	F
1	700000	Cl
1	1000000	Br

[a]Complete list contains 19 entries.

table is suitable for this purpose, a linearized structure is output by means of standard symbols, together with the numbering scheme used.

The numbering scheme is output first by traversing the tree representing the structure in a depth-first algorithm [8]. Then the numbers representing the nodes are replaced by their associated symbols, the nodes being contracted where appropriate.

TABLE 5

Code groups identifying functional groups (octal values)[a]

Central atom		Neighbours		Functionality	Marked neighbours
Part 1	Part 2	Part 1	Part 2		
2	200000	6	100020	Ester —O—	
		Any value	100000		Ester —C—
2	200000	Any value	100000	Ether —O—	
		Any value	100000		
1	200000	Any value	100000	Alcohol —OH	
1	400000	Any value	100000	Amine—NH2	
1	400000	106	100020	Amide —NH2	Carbonyl C
2	400000	106	100020	Amide —NH—	Carbonyl C
		Any value	100000		
6	400020	Any value	100000	Nitro N	
		105	200000		Nitro —O
		0	200001		Nitro =O

[a]Complete list contains 21 entries.

TABLE 6

Perception of functional groups
(The sample structure is the same as that shown in Table 2.)

Atom No.	Code sums[a]		Functionality identified by single code[b]	Functionality identified by code groups[c]	Marked[c]
	Part 1	Part 2			
1	1	100000	Aliphatic C		
2	3	100000	Aliphatic C		
3	2	100020	Ketone C		
4	2	100000	Aliphatic C		
5	6	100020	—	—	Ester =C
6	2	200000	—	Ester —O—	
7	6	100000	Aliphatic C		
8	1	100000	Aliphatic C		
9	1	100000	Aliphatic C		
10	0	200070	Carbonyl O		
11	0	200070	Carbonyl O		

[a]Cf. Table 3. [b]Cf. Table 4. [c]Cf. Table 5.

Chemical shift estimation

For each aliphatic carbon the chemical shift is calculated by adding the increments of the neighbour atoms to the base value. This is done by traversing the tree specified in the connection table with a breadth-first algorithm [8] limited to the δ-level. The increments assigned to the neighbour atoms correspond to the values given earlier [7]. However, to facilitate the computation, increment values given for groups of atoms (e.g. —CO—OC) are split into partial values and assigned to individual atoms in such a way as to reproduce the group increment. Finally, a steric correction, which takes into account the degree of branching of the α-neighbours, is applied where appropriate. Table 7 shows the output resulting from a sample compound together with the measured shift values.

TABLE 7

Output from sample compound together with measured values
Structure: 1 2 [9] 3 [10] 4 5 [11] 6 7 8
CH3—CH[—CH3]—CO—CH2—CO—O—CH2—CH3

Atom No.	Shift	
	Calculated	Measured
1	15.6	18.0
2	45.0	41.2
4	45.0	47.2
7	58.8	61.0
8	13.6	14.2
9	15.6	—

The financial support by the Swiss National Science Foundation and by Spectrospin AG 8117 Zürich-Fällanden is gratefully acknowledged.

REFERENCES

1 V. Formacek, L. Desnoyer, H. P. Kellerhals, T. Keller and J. T. Clerc, ^{13}C Data Bank, Vol. 1, Bruker Physik, Karlsruhe, 1976.
2 L. F. Johnson and W. C. Jankowski, Carbon-13 NMR Spectra, J. Wiley, New York, 1972.
3 E. Pretsch, J. T. Clerc, J. Seibl and W. Simon, Tabellen zur Strukturaufklärung organischer Verbindungen, Springer Verlag, Berlin, 1976.
4 L. M. Jackman and S. Sternhell, Applications of Nuclear Magnetic Resonance Spectroscopy in Organic Chemistry, 2nd edn., Pergamon Press, Oxford, 1969.
5 G. J. Martin, M. L. Martin and S. Odiot, Organic Magnetic Resonance, 7 (1975) 2.
6 J. T. Clerc, E. Pretsch and S. Sternhell, ^{13}C-Kernresonanzspektroskopie, Akademische Verlagsgesellschaft, Frankfurt/Main, 1973.
7 See ref. 3, page C10.
8 N. J. Nilsson, Problem Solving Methods in Artificial Intelligence, McGraw-Hill, New York, 1971.

Analytica Chimica Acta, 95 (1977) 41—49
Computer Techniques and Optimization

AN ON-LINE SEARCH SYSTEM FOR THE MASS SPECTROMETRY LITERATURE

V. A. VINTON and G. W. A. MILNE

Laboratory of Chemistry, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Md., 20014 (U.S.A.)

S. R. HELLER*

Environmental Protection Agency, Washington, D.C., 20460 (U.S.A.)

(Received 21st March 1977)

SUMMARY

Computer programs have been written for interactive, conversational searching through the accumulated files of the 'Mass Spectrometry Bulletin' published by the Mass Spectrometry Data Centre (MSDC). This family of programs is available for use as a part of the NIH/EPA Chemical Information System on a commercial computer network.

The enormous recent increase in the number of groups working on mass spectrometry has led to a variety of problems, amongst which are the difficulties caused by a rapidly expanding literature. Prior to 1960, the annual number of papers dealing with mass spectrometry was small and fairly constant, reflecting the output of the relatively few scientists interested. Then there was a rapid growth of activity, as mass spectrometers that could be used in analytical organic chemistry became available. Further advances have increased this rate of growth so that the number of papers on mass spectrometry is now doubling in less than ten years. In view of this activity, the farsighted decision was made by the MSDC in 1965 to examine the literature on a day-to-day basis for papers dealing with mass spectrometry. Citations and abstracts were collected and published on a monthly basis as the Mass Spectrometry Bulletin [1], which is now a generally recognized authority on the literature of mass spectrometry.

As with any printed data base, the Bulletin becomes more difficult to use as it gets larger. Apart from the physical size of the accumulated material, indexing becomes more crucial and cumulative indexes are needed. Simultaneously, the value of computer methods for handling these problems has become clearer. Since 1970, the U.K. and U.S. Governments have collaborated, in the development of the MSDC-NIH-EPA Mass Spectral Search System (MSSS) [2]. As an extension of this work, a study of the use of the computer-readable files of the Bulletin as the basis of an on-line, interactive literature search system has been undertaken. This system is the subject of this paper.

Data base

The Bulletin currently publishes some 6,500 citations per year. The complete data base now requires about 20 million characters of storage space. By the end of 1975 there were a total of 56 567 citations; 139 392 author entries involving 49 309 authors and 42 632 element entries. There are 272 741 entries in the subject index, 46 735 in the general index, and 76 520 compounds are identified by compound class.

SEARCH PROGRAMS

In 1970, Hertz et al. [3] described a computer search system that used the files of the Mass Spectrometry Bulletin. This system was designed to use tape copies of the data base and so overcome the difficulties associated with a search system operating on a central large computer. The programs were run on an IBM 1802 and permitted searches for papers according to author, subject or element. Alternatively, searches could be made for compound type, elemental composition or molecular weight. Intersected searches, e.g. for all indoles with a given molecular weight, could be effected. These programs used all the parameters that had been used in the indexing of the Bulletin and so offered an alternative means of using the information in the Bulletin. They also allowed, in principle, an individual to conduct the search without recourse to a distant computer. The single disadvantage of the system is that because the search is of a data base that is stored on magnetic tape, it is relatively slow; a typical search requires about 5 min, the limiting step being the transfer of data from the tape.

Since 1971, the use of disk, as opposed to tape, for the storage of large, searchable data bases has been examined. The cost of disk storage for a given data base, while considerably more than the cost of tape storage for the same data, is currently decreasing very rapidly, and now the cost of storing even a large data base on disk can be manageable. The great advantage of disk over tape is the vastly greater speed with which the information can be located and retrieved. This rapid retrieval permits construction of a truly interactive search system, in which the user, having asked a question, need wait only moments to receive an answer which may be sufficient or may permit re-statement of the question.

The NIH-EPA Chemical Information System (CIS), a collection of numerical and literature data bases together with the necessary search programs [4], uses disk-stored data bases exclusively. Much of the CIS runs on a commercial networked computer which provides on-line access to data — a feature which demands that the data be stored on disk. Each data base is organized into two or three separate disk files, as has been described for the MSSS [5]. The major one is the so-called reference file which contains all the data necessary for a search, organized in a specific way. The other file is a pointer file, which is the file that is first reached by the search program. From the pointer file, the program finds where the information needed begins in the reference file

and where it ends. The program then reads that part of the reference file in preparing the answer to the user's query.

As an example, if the user asks for all papers by Atkins to be located, the program goes to the appropriate pointer file, which is simply an alphabetically ordered file of the names of all authors whose papers are in the data base. Once the name 'Atkins' has been found in the pointer file, it is simple to read off starting and ending addresses of 'Atkins' entries in the reference file, and from these two numbers, the number of 'Atkins' entries can be calculated and reported back to the user.

With this general technique, the data base can be searched in a variety of ways. Every paper, when it is taken from the literature, is assigned a certain number of subject keywords from a list that currently has 327 entries. In addition, any element that is studied specifically is noted as are the names of all the authors and the MSDC Compound Classification Codes of compounds described. The MSDC codes are a group of 87 four-digit codes that represent functional groups or compound types appropriate to the compounds discussed. In addition to these descriptors, there is also a 'general index' of terms which do not logically belong in the other indexes. This index includes, for example, the names of microorganisms that have been studied and the venues of meetings dealing with mass spectrometry.

These five descriptors, subject, element, author, compound class and general index term, form the basis of the indexes to the Bulletin; they can readily be stripped from the data base as it is written on the magnetic tape and they have therefore been used as the parameters with which the computer search can operate.

RESULTS

In the subject search, the user is asked to enter the subject of interest. To counter ambiguities in phraseology and spelling, the system searches through the data base for occurrence of the first three letters provided by the user, so if the word 'electron' is entered, entries to subject codes such as electron, electric and electronic will be retrieved. These are then presented to the user who is asked to select one or more of these temporary files. The program then accepts another subject word or, if so requested, lists the references that have already been retrieved. Before any references are listed, however, the user is asked whether or not references should be listed irrespective of the year of their appearance in the Bulletin. An affirmative answer results in all the references being printed, but a negative answer requires the user to specify the years for which citations should be retrieved. Once the years in question, (e.g. 73, 74, 75), are defined, only those references will be listed.

An example of such a search is given in Table 1. Here the user was searching for any papers dealing with mass measurement in connection with chemical ionization mass spectrometry. The first subject word entered was 'mass measurement' and the subjects that were encountered were mass

TABLE 1

Subject search in the Mass Spectrometry Bulletin

SUBJECT SEARCH

TYPE FIRST 3 LETTERS OF SUBJECT NAME
CR TO EXIT, 1 FOR LIST OF REFERENCES

SUBJECT: MASS MEASUREMENT

THE FILE CONTAINS

INDEX #	# OF REFS	SUBJECT
1	194	MASS DISCRIMINATION
2	266	MASS FRAGMENTATION
3	463	MASS MEASUREMENT
4	11422	MASS SPECTRA
5	817	SPARK SOURCE MASS SPECTROMETRY
6	375	THEORY OF MASS SPECTRA
7	357	TIME RESOLVED MASS SPECTROSCOPY

ENTER THE INDEX NUMBER (S) CORRESPONDING TO THE NAME (S) YOU WANT
INDEX NUMBERS: 3

# REFS	SUBJECTS
463	MASS MEASUREMENT

NEXT SUBJECT: CHEMICAL IONIZATION

THE FILE CONTAINS

INDEX #	# OF REFS	SUBJECT
1	356	CHEMICAL BINDING ENERGY
2	678	CHEMICAL IONIZATION
3	1584	CHEMICAL REACTIONS
4	608	RADIATION CHEMISTRY

ENTER THE INDEX NUMBER (S) CORRESPONDING TO THE NAME (S) YOU WANT
INDEX NUMBERS:2

# REFS	SUBJECTS
8	MASS MEASUREMENT
	CHEMICAL IONIZATION

discrimination (file 1, 194 references), mass fragmentation (266 references), mass measurement (463), mass spectra (11 422), spark-source mass spectrometry (817), theory of mass spectra (375), and time-resolved mass spectrometry (357) — all subject phrases containing the word 'mass'. The user is then asked which of the seven answer files are of interest and responds by specifying file 3.

The other files are then discarded and the user is prompted for a second subject word. This is 'chemical ionization', which prompts a search which leads to answer files for 'chemical binding energy', 'chemical ionization', 'chemical reactions', and 'radiation chemistry'. Once the user selects the second of these, chemical ionization, these 678 references are automatically combined in a Boolean AND operation with the 463 papers previously retrieved as relevant to mass measurement. The result is an answer file containing the 8 papers that discuss mass measurement and chemical-ionization mass spectrometry. These can be listed for any or all Bulletin years, or a third subject term can be introduced and the search continued. This complete session, from log-in to log-off takes less than 30 s and the DEC PDP-10 cpu time required is less than 5 s.

There are currently a total of 327 specific subject code words in the data base. These range from the relatively non-specific (e.g. impact, electric and so on), to quite precise terms such as scatter-electrons or quadrupole.

While the example in Table 1 illustrates the use of a Boolean AND combination of two separate queries, the use of a NOT combination is also possible. If the user specifies a subject word, followed by the operator NOT, then the program will produce all the citations that do not deal with that particular subject.

In the preparation of the Bulletin, the abstracts of papers that deal with specific elements are so coded. Elements that are merely part of a molecule under study (e.g. C, H, O, N etc.) are not usually noted, but a paper dealing with the secondary ion emission from copper surfaces, for example, would be considered as dealing with copper, and copper would be one of the keywords. The element search simply uses these element keywords to retrieve all the papers that deal with specific elements. In using this program, one has simply to specify the atomic symbol of the element in question and the program, using the pointer file/reference file technique discussed earlier, responds by telling the user how many citations are retrieved. The user is then free to list these (limiting the listing to certain years if so desired) or enter another element and continue to narrow down the number of citations. As in the subject search, AND or NOT combinations are possible. Thus a search for papers dealing with iron produced 483 references (Table 2). Only 146 papers, however, deal with both iron and cobalt, and if those papers that also deal with nickel are excluded, 24 references remain. This program is of particular value to those interested in inorganic chemistry, metallurgy and surface phenomena, and demonstrates the breadth of coverage of the Bulletin beyond the boundaries of organic mass spectrometry.

A third type of search is for papers published by a specific author or group of authors. A particular problem here is that people's surnames are, in effect, trivial names and are not spelled predictably. Cursory examination of the Bulletin file shows that at some step of the publication and abstracting process, a given author's name may be changed. The author search developed for this system attempts to deal with this problem by prompting the user for

TABLE 2

Element search in the Mass Spectrometry Bulletin

ELEMENT SEARCH

TYPE ELEMENT SYMBOL

CR TO EXIT, 1 FOR LIST OF REFERENCES

ELEMENT: FE

# REFS	ELEMENTS
483	FE

NEXT ELEMENT: CO

# REFS	ELEMENTS	
146	FE	CO

NEXT ELEMENT: NI NOT

# REFS	ELEMENTS		
24	FE	CO	—NI

the name of the author whose papers are the object of the search. The actual search however, is carried out with the first four letters of the surname. All the papers by authors whose surnames begin with these four letters will be retrieved and the user must then decide which are appropriate. Table 3 gives an example of an author search in which the goal was to locate any papers published by Aberth and Anbar of Stanford Research Institute. When the first name, Aberth, is entered, a list of nine authors whose names begin with 'aber' is produced. From this list, it can be seen that the same name may be spelled in more than one way and people's initials are far from constant. In this case, it seems reasonable to guess that W. Aberth and W. H. Aberth are the same person, and so index numbers 8 and 9 are selected, resulting in a file of 36 papers by one or other of those authors. When the next author, Anbar, is specified, the result is simpler because he is unique in having 'anba' as the first four letters of his name. When the two files — Aberth papers and Anbar papers — are intersected, 10 papers by Anbar and either W. Aberth or W. H. Aberth are present. The observation that W. Aberth and W. H. Aberth both publish with Anbar, but never together strongly suggests that they must be the same person.

As in the searches discussed above, NOT logic is available in the author search. For example, one could, by using the operator NOT after Anbar's name, retrieve all the papers that the Aberths have published without Anbar as co-author.

The use of the first four letters of an author's name as described here has the obvious advantage that the user need not spell the full name accurately and, significantly, need not know the correct initials. Even if the user knows

TABLE 3

Author search in the Mass Spectrometry Bulletin

```
AUTHOR SEARCH

TYPE FIRST 4 LETTERS OF AUTHOR'S NAME
CR TO EXIT, 1 FOR LIST OF REFERENCES

AUTHOR:  ABER

THE FILE CONTAINS
INDEX #   # OF REFS      AUTHOR
  1           1         ABERCROMBIE I.D.
  2           1         ABERG G.
  3           4         ABERG T.
  4           6         ABERHART D.J.
  5           1         ABERHART J.
  6           8         ABERNATHEY R.M.
  7           1         ABERNATHY R.M.
  8          20         ABERTH W.
  9          16         ABERTH W.H.

ENTER THE INDEX NUMBERS (S) CORRESPONDING TO THE NAME (S) YOU WANT
INDEX NUMBERS: 8 9
     # REFS      AUTHORS

        36       ABERTH W.
                 ABERTH W.H.

NEXT AUTHOR:  ANBAR

THE FILE CONTAINS
INDEX #      # OF REFS        AUTHOR
  1             21          ANBAR M.

ENTER THE INDEX NUMBER (S) CORRESPONDING TO THE NAME(S) YOU WANT
INDEX NUMBERS: 1
     # REFS      AUTHORS

        10       ABERTH W.
                 ABERTH W.H.
                 ANBAR M.
```

the correct initials, the search need not be accurate because, as the example in Table 3 shows, scientists do not use their own initials consistently. The major disadvantage of the four-letter search is that an entry such as 'smit' produces a total of 167 candidates, including 159 Smiths. Use of a five-letter search would not solve this problem, and it may prove necessary to permit the user to supply author's initials to alleviate this difficulty. The only inconvenience of the search as it stands however, is the time it takes to list all the candidates; this listing does not affect the computer costs seriously.

The published version of the Bulletin contains a 'General Index', which in earlier issues was termed the 'Materials Index'. This index contains entries which cannot sensibly be placed in other indexes and yet are useful for information retrieval. Such entries include types of ions (e.g. Ar(+)) and compounds (e.g. metal chelates), biological classifications (e.g. *Klebsiella*) and even the venues (Berlin, Dallas) of mass spectrometry meetings. As a result,

TABLE 4

General Index search in the Mass Spectrometry Bulletin

```
GENERAL INDEX SEARCH
TYPE FIRST 4 LETTERS OF GENERAL INDEX ITEM
CR TO EXIT, 1 FOR LIST OF REFERENCES

ITEM: CO2

THE FILE CONTAINS
INDEX #      # OF REFS          ITEM
  1            532            CO2
  2              1            CO2 GAS
  3              1            CO2 LIQUID

ENTER THE INDEX NUMBER(S) CORRESPONDING TO THE ITEM(S) YOU WANT
INDEX NUMBERS: 1
     # REFS        ITEMS
       532       CO2

NEXT ITEM: BIARR

THE FILE CONTAINS
INDEX #      # OF REFS          ITEM
  1            35             BIARRITZ

ENTER THE INDEX NUMBER(S) CORRESPONDING TO THE ITEM(S) YOU WANT
INDEX NUMBERS: 1
     # REFS        ITEMS
       1         CO2
                 BIARRITZ

NEXT ITEM: 1

ALL BULLETIN YEARS? (Y OR N)

Y

 906523   LUMINESCENCE FROM LOW ENERGY ION MOLECULE INTERACTION MA
          RX R. BOOK. INTERACTION BETWEEN IONS AND MOLECULES P.
          AUSLOOS ED., PLENUM PRESS, NEW YORK AND LONDON.  NATO ADV
          ANCED STUDY INST., BIARRITZ, FRANCE  24 JUNE-6 JULY 1974
          P.563-77 1975.
```

this General Index can occasionally be extraordinarily useful as a means of locating a specific piece of information.

The program 'Index' can be used to search through the accumulated general indexes and operates in the same way as the subject search described above. An example of the use of this program is given in Table 4. In the first case, the user was able very rapidly to locate the one paper dealing with carbon dioxide that was presented at the 1975 NATO Advanced Study Institute held in Biarritz.

This program, like the author search, uses the first four letters of the index term supplied by the user. This rarely produces more than a few possible answers; problems such as that arising from names like 'Smith' apparently do not occur in this case. The General Index is not static and can grow more or less rapidly as more papers are abstracted. This raises the possibility that newly introduced terms cannot be used immediately for searching. Such terms will be inaccessible until the next annual update of the searchable data base.

Dissemination of the system

The Mass Spectrometry Search System (MSBULL) is available as an extension of the Mass Spectral Search System. This system is maintained by MSDC on the ADP-Cyphernetics international computer network. In addition to the searching programs, the system, like MSSS, contains user assistance files that can be accessed by the user who needs help with the conversational style of the system. The user can also list the currently used subject codewords or the MSDC compound classification codes.

We acknowledge the generous cooperation received from Dr. A. McCormick and Mr. D. C. Maxwell of the Mass Spectrometry Data Centre, Aldermaston, England, and from Mr. A. C. Nicholas of the Department of Industry of the U.K. Government.

REFERENCES

1 Mass Spectrometry Bulletin, Mass Spectrometry Data Centre, AWRE, Aldermaston, Berks., England.
2 S. R. Heller, H. M. Fales and G. W. A. Milne, Org. Mass Spectrom., 7 (1973) 107; S. R. Heller, D. A. Koniver, H. M. Fales and G. W. A. Milne, Anal. Chem., 46, 947 (1974); S. R. Heller, R. J. Feldmann, H. M. Fales, and G. W. A. Milne, J. Chem. Doc., 13 (1973) 130; G. W. A. Milne and S. R. Heller, J. Chem. Inf. Comp. Sci., (1976) in press.
3 H. S. Hertz, D. A. Evans and K. Biemann, Org. Mass Spectrom., 4 (1970) 453.
4 S. R. Heller, G. W. A. Milne and R. J. Feldmann, Science, (1977) in press.
5 S. R. Heller, Anal. Chem., 44 (1972) 1951.

Analytica Chimica Acta, 95 (1977) 51—58
Computer Techniques and Optimization

ANALYSIS OF BINARY MIXTURES BY COMPUTER DECOMPOSITION OF MOLECULAR FLUORESCENCE SPECTRA

CARL E. RECHSTEINER, Jr., HARVEY S. GOLD, and RICHARD P. BUCK*

William R. Kenan Laboratories of Chemistry, University of North Carolina, Chapel Hill, NC 27514 (U.S.A.)

(Received 18th January 1977)

SUMMARY

Qualitative analysis by computer decomposition of fluorescence spectra by means of the program SPECSOLV is discussed. Concentration studies of naphthacene in benzene and mixture analyses of single and binary solutions of anthracene, naphthacene, naphthalene, and rubrene are reported. The results demonstrate the ability to separate solvent and sample spectral contributions, and to assign component peaks in the case of mixtures. Semi-quantitative results are presented, and the feasibility of extending the study to computer-search systems based on component characterization of fluorescence spectra is discussed.

The spectrometric identification of single-component materials has become accepted analytical practice, but the routine identification and determination of multicomponent mixtures from spectra has remained an elusive goal. A recent article [1] discusses the theoretical relationship between the fluorescence spectra of two- and three-component mixtures and the spectra of individual components. However, the experimental quantification of multicomponent mixtures is limited by special conditions on peak overlap and normalized intensity [2]. The nature of certain biological systems sometimes permits accurate analysis [3—5]. Other approaches have included fluorescence correlation spectroscopy [6] and electron-beam fluorescence [7]. A concurrent, but previously unrelated, area of research has involved computer decomposition of spectra to obtain peak parameters for individual components. In the current study, computer decomposition methods have been used to achieve identification of mixtures of materials by their molecular fluorescence.

Decomposition of spectra by curve-fitting methods has been widely applied to many spectral types. Such areas as infrared spectroscopy [8—13] have been studied in great detail, as have u.v.—visible absorption spectroscopy and chromatography. However, emission processes such as molecular fluorescence have been largely overlooked. A generalized spectral decomposition method which is applicable to a wide variety of spectra has recently been reported [14]. Miller and Faulkner [15] have shown that characterized fluorescence spectra of single compounds can be successfully matched against a computer file of similarly characterized reference materials.

The technique of Miller and Faulkner is significant; it demonstrates that fluorescence spectral characteristics can be used similarly to those of infrared spectroscopy [16—21] and mass spectrometry [22—28]. The method used, however, depends on a first-derivative search for peaks in the fluorescence spectra. Peak positions and relative intensities are then used in the identification search scheme. While not considered by Miller and Faulkner, the practical problems of locating and characterizing peaks in multicomponent mixtures are clearly far more difficult. This is due to the inherent features of the fluorescence process, which gives rise to typically broad and overlapping peaks, whose relative and absolute intensities depend on variables such as excitation frequency, solvent polarity, and concentration.

The broad, not easily defined, peaks characteristic of molecular fluorescence do not lend themselves to easy correlation with molecular structure or to quantitative work when bands overlap. Nevertheless, the information content of these spectra is quite high. The key to extracting this information involves accurate determination of individual peak parameters. To this end, a computer decomposition method based on iterative solution and least-squares fit of component bands is ideally suited.

With this approach, peak parameters characteristic of a given compound are easily and readily obtained. These parameters of peak location, peak width descriptors, and intensities are easily obtained by use of a FORTRAN computer program, i.e., SPECSOLV, [14, 29]. Fluorescence peaks are generally ideally additive when multicomponent mixtures are analyzed. It is of major significance that SPECSOLV is able to decompose the fluorescence spectrum of a mixture into peaks which can be easily and generally unambiguously identified as those present in the spectra of the individual components. Even without considering such factors as quenching and concentration effects, semi-quantitative results are readily obtainable.

EXPERIMENTAL

Instrumentation

A Hitachi MPF-2A fluorescence spectrometer was used in the direct mode to obtain the emission spectra of the test compounds. Either a 274-nm excitation (for the concentration studies) or a 290-nm excitation (for the mixture analysis) was used with 10-nm emission and excitation slits. Digital data were recorded from a Keithley model 160 Digital Multimeter.

Reagents

Spectro-quality cyclohexane (MC/B) and A.C.S. reagent-grade benzene were used as solvents. Anthracene (Aldrich Gold Label, 99.9%), naphthacene, naphthalene, and rubrene (Baker) were quantitatively dissolved in the appropriate solvent with ultrasonic shaking.

Data analysis

The resulting digitized spectrofluorimetric data were analyzed by the SPECSOLV program, which is a catalogued procedure available at both the Triangle Universities and University of North Carolina Computation Centers [29].

RESULTS

The application of SPECSOLV to various spectroscopic problems has been shown elsewhere [14]. It can be applied to spectrofluorescence because a broad, slowly changing spectrum can be converted to a set of readily accessible numerical characterization parameters. Implicit is the assumption that the spectral components can be adequately described by a Gaussian or by an asymmetric Gaussian ("bi-Gaussian") curve. In 1930, Kuhn and Braun [30] proposed a formulation for ultraviolet—visible absorption spectra in terms of a Gaussian function in wavenumber. The similar nature of the excited states for molecular absorption and emission allows a parallel derivation for molecular fluorescence, and therefore permits the current assumption of Gaussian or bi-Gaussian form. Clearly, the Gaussian form will be evident only for spectra linear in wavenumber or other energy units. For ease of use, SPECSOLV accepts molecular fluorescence spectra in Angstrom units, which are then internally transformed to kilokaysers (1.0 kayser = 1.0 cm^{-1}), and all peak characterization is done in these units. Again for convenience, results and all spectral plots are presented in Angstroms. The characterization parameters can be rapidly compared with reference data sets (generated from known compounds by parallel analysis) to determine the identity and composition of fluorescent mixtures.

Concentration studies

To be utilized to maximum advantage, a method of analysis based on decomposition must identify components over a wide range of concentrations. Solvent peaks must be identified so that these can be eliminated from consideration before any attempt at comparison with standard reference materials. Each of the individual peaks of a spectrum is characterized by four numbers representing the intensity, wavelength of peak maximum, and two shape parameters that describe the right and left portions of each peak relative to the peak maximum. The emission peaks for several concentrations of naphthacene dissolved in benzene are shown in Table 1, with the peaks ranked according to relative intensity. Peak positions for the benzene solvent and for naphthacene dissolved in cyclohexane are also presented and identified.

The procedure gives rise to a relative error of ±1 nm in the instrument wavelength setting. Coupled with the 10-nm slit widths, this leads to an expected error of ±3 nm in the position of peak maxima for each spectrum. Nevertheless, there is excellent agreement for the peak positions over three orders of magnitude in naphthacene concentration. The reversal in the ranking of the peak at 344.0 nm at high concentrations is probably due either to quenching of that transition or to self-absorption, since both processes are

TABLE 1

Computed peak wavelength maxima (nm) for naphthacene dissolved in benzene at various concentrations.[a]

Benzene	490	49	4.9	0.49	11.0[b]
306.3	479.6	344.2	306.2[c]	305.8[c]	473.4
329.9	510.9	477.9	343.3	323.4[c]	510.9
386.7	344.7	510.7	476.6	283.0[c]	344.7
360.1	546.5	547.5	328.6[c]	350.2[d]	541.0
—	339.8	339.9	511.3	477.6	—
—	—	306.9[c]	286.6[c]	511.9	—

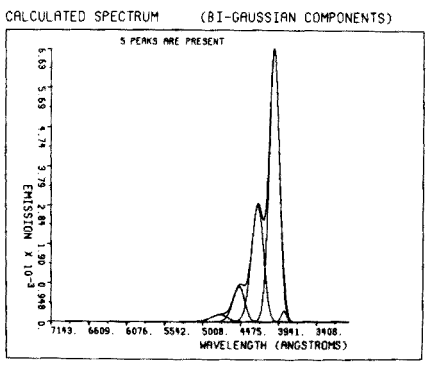
[a]Peaks with an intensity greater than or equal to 10% of the maximum peak intensity are listed.
[b]Dissolved in cyclohexane; 290-nm excitation.
[c]Benzene component peaks.
[d]350.2-nm peak shifted from 344 nm owing to differences in computational limits.

induced by high concentration. Further evidence for this is the extremely close match of the peak-shape parameters for that peak in all of the naphthacene-containing samples (about 2% variation for those samples dissolved in benzene). Peaks caused by benzene emission can be easily eliminated by reference to comparison spectra.

Mixture analysis

The experiments discussed above demonstrate that solvent and sample fluorescence can be distinguished, and imply that multicomponent mixture spectra can be easily dealt with. To approach this problem, 290-nm excitation was used to study single- and binary-component mixtures of anthracene, rubrene, naphthalene, and naphthacene in cyclohexane. These mixtures are classic examples of systems which exhibit severe overlap of component peaks. The 290-nm excitation causes fluorescence in each compound, but is not optimized for any one of them. Figures 1—7 show the emission spectra at 290-nm excitation for each of the standard compounds, and for three of the six possible binary mixtures. To reduce the computational requirements for this study, a reasonable limit of five peaks for each individual spectrum and ten peaks for each of the binary mixtures (anthracene—naphthalene, naphthacene—naphthalene, and naphthacene—rubrene) was assumed. This choice proved to be valid as indicated by program convergence in each analysis. The set of selected binary mixtures contains cases of widely spaced, moderately spaced, and closely spaced spectral overlaps which lead to an easy assessment of the resolution achievable through use of SPECSOLV.

Table 2 lists representative peak parameters for the individual standard compounds (anthracene, naphthacene, naphthalene, and rubrene). Table 3 lists and assigns the seven most prominent peaks determined by data analysis of the mixture spectra. These results clearly demonstrate the feasibility of this approach, even when two fluorescent compounds exhibit severe overlap. In point, naphthacene and naphthalene each contain components with re-



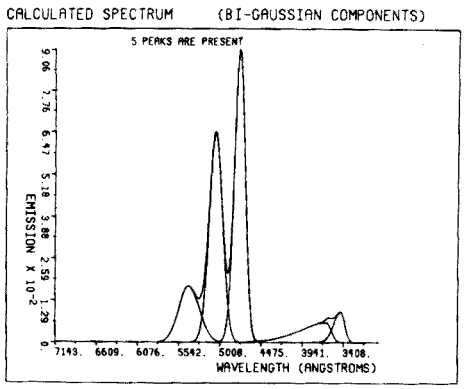
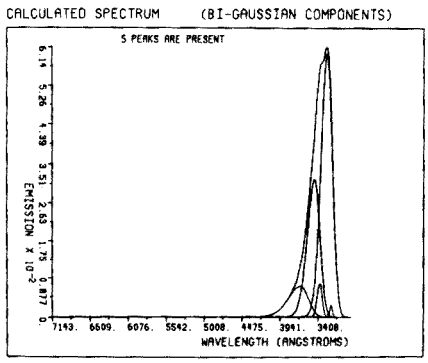


Fig. 1. Anthracene input spectrum with component peaks.

Fig. 2. Naphthacene input spectrum with component peaks.



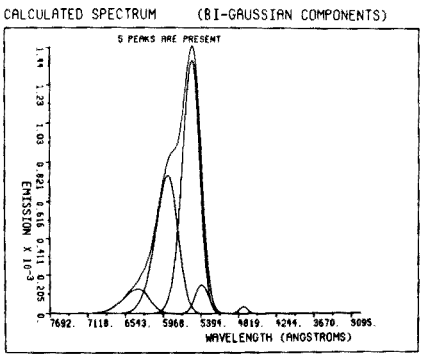
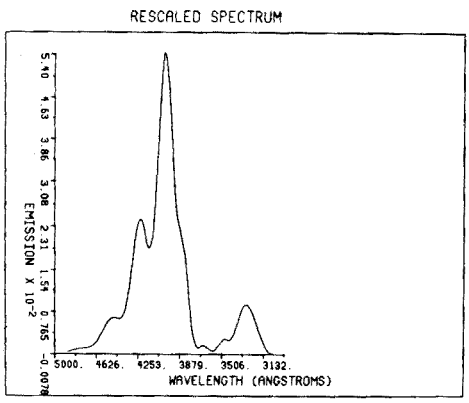


Fig. 3. Naphthalene input spectrum with component peaks.

Fig. 4. Rubrene input spectrum with component peaks.



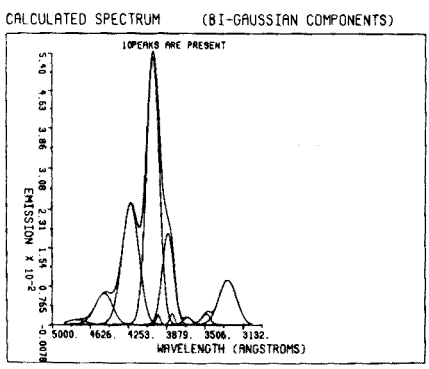


Fig. 5. (A) (left) Anthracene—naphthalene binary mixture spectrum. (B) (right) Component peaks. Peaks with maxima at 364.5, 379.8, and 393.8 nm are attributed to noise.

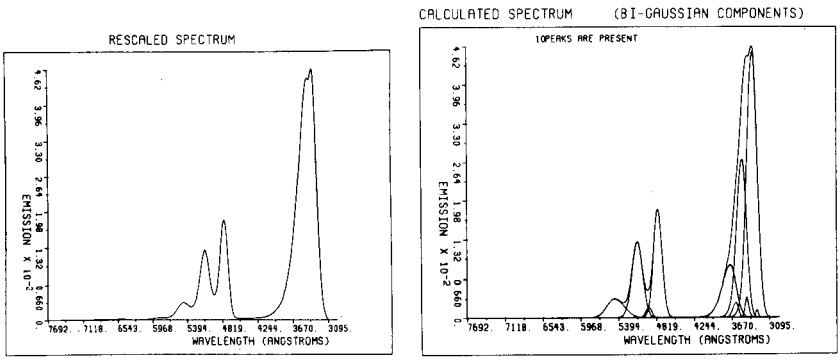


Fig. 6. (A) (left) Naphthacene—naphthalene binary mixture spectrum. (B) (right) Component peaks. Peaks with maxima at 310.5, 352.0, and 487.7 nm are attributed to noise.

TABLE 2

Component peak parameters for pure reference materials

	Wavelength (nm)	Intensity	Peak parameter (+)	Peak parameter (—)
Anthracene	399.1	6.60	8.82	7.99
(85 μg ml^{-1})	422.1	2.80	10.05	10.01
	447.8	0.86	9.15	10.23
	384.3	0.27	4.21	4.21
	475.1	0.17	12.91	14.33
Naphthacene	473.3	9.10	7.79	8.38
(11 μg ml^{-1})	505.0	6.40	10.51	9.64
	541.0	1.70	17.80	13.59
	341.9	0.93	6.49	10.14
	360.3	0.61	7.69	36.32
Naphthalene	325.0	6.00	8.79	9.89
(136 μg ml^{-1})	342.4	3.20	8.54	11.87
	363.0	0.68	13.62	19.07
	319.6	0.26	2.28	2.28
	335.4	0.75	3.76	4.46
Rubrene	551.2	1.37	14.21	16.89
(33 μg ml^{-1})	588.0	0.75	18.14	21.38
	633.4	0.13	20.03	25.66
	472.3	0.03	6.64	7.43
	537.2	0.15	13.43	10.53

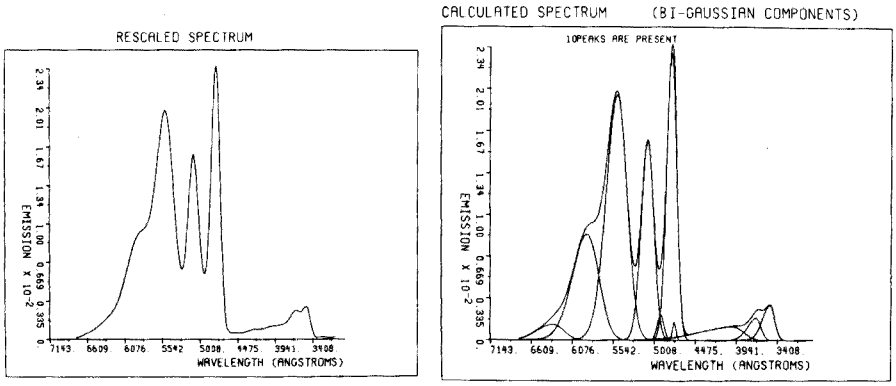


Fig. 7. (A) (left) Naphthacene—rubrene binary mixture spectrum. (B) (right) Component peaks. Peaks with maxima at 392.5, 487.3, and 469.2 nm are attributed to noise.

TABLE 3

Component peak parameters for binary mixtures

	Wavelength (nm)	Intensity	Parameter (+)	Parameter (—)	Contributing compound label[a]
Naphthalene/ Anthracene	399.4	5.33	7.59	7.22	A
	421.6	2.42	10.00	9.52	A
	384.3	1.84	7.02	6.42	A
	325.0	0.89	11.04	9.49	Nl
	447.5	0.62	9.72	10.43	A
	345.6	0.21	4.43	5.98	Nl
	475.3	0.09	10.68	10.68	A
Naphthalene/ Naphthacene	325.2	4.52	8.78	9.99	Nl
	341.2	2.69	7.49	10.19	Nl
	473.0	1.83	7.62	7.94	Nc
	504.9	1.27	9.49	9.71	Nc
	360.4	0.89	10.54	14.66	Nl
	541.3	0.33	18.02	13.04	Nc
	334.7	0.35	3.26	3.26	Nl
Naphthacene/ Rubrene	473.3	2.31	7.60	7.80	Nc
	545.7	1.96	15.77	16.53	R
	505.1	1.59	9.81	9.71	Nc
	585.7	0.85	19.42	21.41	R
	342.5	0.28	7.21	11.19	Nc
	360.9	0.18	7.65	12.65	Nc
	630.5	0.13	17.39	23.22	R

[a] A = anthracene. Nl = naphthalene. Nc = naphthacene. R = rubrene.

spective peak maxima within 4 nm of one another, yet the peak parameters allow classification of the peak at 341.2 nm as being due to naphthalene rather than naphthacene. The ability to assign component peaks unambiguously to a particular compound, even in the case of severe overlap, leads directly to a new capability of handling mixture spectra conveniently in a straightforward manner.

This study has demonstrated the ability of SPECSOLV to characterize pure materials as well as to aid in identification of components of mixtures over a wide range of sample concentrations. Extension of SPECSOLV to search systems for fluorescence data follows directly and should be particularly useful for low-level determinations. Further work is in progress to complete the extension of this method to the systematic development of computer searchable library files of spectrofluorescence data, and to enhance the quantitative accuracy of this method for low concentrations.

The authors thank M. E. Scott for assistance in the early stages of this work, which was supported by a grant from the National Science Foundation, MPS75-00970.

REFERENCES

1 I. Ketskemety and J. Kusba, Acta Phys. Chem., 20 (1974) 239.
2 T. L. Pasby, in A. J. Pesce (Ed.), Fluorescence Spectroscopy, M. Dekker, New York, 1971, 149—201.
3 P. Einersson, H. Hallman, and G. Jonsson, Med. Biol., 53 (1975) 15.
4 O. Lindvall, A. Bjorklund, and B. Talch, J. Histochem. Cytochem., 23 (1975) 703.
5 V. A. Gordyskii and A. A. Tikhomolov, Opt. Spektrosk., 38 (1975) 875.
6 S. R. Aragon and R. Pecora, J. Chem. Phys., 64 (1976) 1791.
7 P. V. C. Hough, W. R. McKinney, M. C. Ledbetter, R. E. Pollack, and H. W. Moos, Proc. Nat. Acad. Sci. U.S.A., 73 (1976) 317.
8 R. D. Fraser and E. Suzuki, Anal. Chem., 41 (1969) 37.
9 K. S. Seshadri and R. N. Jones, Spectrochim. Acta, Part A, 19 (1963) 1013.
10 R. P. Young and R. N. Jones, Chem. Rev., 71 (1971) 219.
11 A. M. Kabiel and C. H. Boutros, Appl. Spectrosc., 22 (1968) 121.
12 F. C. Strong, Appl. Spectrosc., 23 (1963) 593.
13 J. Pitha and R. N. Jones, Can. J. Chem., 44 (1966) 3031.
14 H. S. Gold, C. E. Rechsteiner, and R. P. Buck, Anal. Chem., 48 (1976) 1540.
15 T. C. Miller and L. P. Faulkner, Anal. Chem., 48 (1976) 2083.
16 D. S. Erley, Anal. Chem., 40 (1968) 894.
17 F. E. Lytle, Anal. Chem., 42 (1970) 355.
18 D. S. Erley, Appl. Spectrosc., 25 (1971) 200.
19 R. W. Sebesta and G. G. Johnson, Jr., Anal. Chem., 44 (1972) 260.
20 E. C. Penski, D. A. Padowski, and J. B. Bouch, Anal. Chem., 46 (1974) 955.
21 R. C. Fox, Anal. Chem., 48 (1976) 717.
22 T. O. Gronneberg, N. A. B. Gray, and G. Ellington, Anal. Chem., 47 (1975) 415.
23 S. R. Heller, D. A. Koniver, H. M. Fales, and G. W. A. Milne, Anal. Chem., 46 (1974) 947.
24 S. L. Grotch, Anal. Chem., 45 (1973) 2.
25 S. R. Heller, Anal. Chem., 44 (1972) 1951.
26 S. L. Grotch, Anal. Chem., 43 (1971) 1362.
27 B. A. Knoch, I. C. Smith, D. E. Wright, R. G. Ridley, and W. Kelly, Anal. Chem., 42 (1970) 1516.
28 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1464.
29 H. S. Gold, SPECSOLV — A generalized spectral decomposition program, Library Service Series Document No. LS-301-0, Research Triangle Park (N.C.), Triangle Universities Computation Center, 1976.
30 W. Kuhn and E. Braun, Z. Phys. Chem., Abt. B, 8 (1930) 281; 9 (1930) 426.

Analytica Chimica Acta, 95 (1977) 59—73
Computer Techniques and Optimization

ALGORITHMS FOR MODELING AND PROCESSING SPATIAL INFORMATION IN HETEROGENEOUS PLASMA DISCHARGES

ALEXANDER SCHEELINE and JOHN P. WALTERS*

Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706 (U.S.A.)

(Received 20th April, 1977)

SUMMARY

Procedures are described for modeling and analyzing optical phenomena in bilaterally symmetric plasma discharges. Discharge shape, emission, absorption, background subtraction, and spatial instability are considered, and provision is made for imperfect imaging properties of the observation system. Algorithms are presented for interconverting lateral, laboratory information concerning a plasma discharge and radial, chemical information concerning the discharge. Various non-idealities and non-emissive processes may be modeled. The assumptions made are that the discharge is cylindrically symmetric and is viewed in a manner which preserves this symmetry; that the optical system non-ideality is limited to depth-of-field distortions, which can be corrected; and that emission and absorption take place in concentric but segregated regions of the discharge, with emission occurring closer to the discharge core. There are known limitations to this model, and any seemingly inconsistent results in experimental applications will indicate that processes not accounted for are active.

Spatial resolution of the numerous properties of spectral light sources may give insights into the mechanisms of operation of these sources and thence improved analytical procedures [1—2]. The spatially heterogeneous structure of d.c. arcs [3—8], inductively coupled plasmas [9, 10], and stabilized spark discharges [11, 12] have been studied. Phenomena other than emission in radiant light sources must be considered in order to determine adequately the spatial distribution of emitting species within a discharge [13]. The present paper is devoted to outlining the algorithms used in generating the plots presented earlier [13]. The order of presentation is similar to that in the earlier paper, to facilitate comparison of results with computational methods.

Abel Inversion is the major technique employed. The Inversion interconverts cylindrically symmetric, radially inhomogeneous properties of a discharge and laterally inhomogeneous observable profiles, by means of matrices derived from the geometry of the discharge and observation system. The mathematics involves solution of a set of simultaneous equations, with the number of equations equaling the number of discrete spatial regions in the discharge. The setup of the equations has been explained [3, 11, 13].

COMPUTATION OF AREA MATRICES

A (cylindrical) plasma discharge may be spatially resolved in a number of dimensions (see Fig. 1). One convenient dissection of the discharge is observation of a single axial slice, which is viewed experimentally in a series of lateral zones [13]. The lateral information may be converted to radial information by means of the Abel Inversion described earlier [13] and in eqns. (5) and (6). The area coefficients needed to construct the conversion matrices are computed by obtaining the area of overlap between individual radial rings and lateral zones. These areas were determined by using closed-form expressions for the integrals involved. In addition, the procedure described below allows correction of the area matrices for the effects of optical depth-of-field. Discrete depth-of-field regions are defined perpendicular to the optical axis and of thickness equal to the width of the lateral zones, and weighting factors are used to "distort" the calculated areas to conform to those viewed by the non-ideal optical observation system.

For simplicity, the calculations will be described in two stages: (1) the procedure for computing areas without accounting for the depth-of-field weighting factors; and (2) the method for including such weighting. The area of overlap between lateral zone i and radial ring j is symbolized as A_{ij}. Explicitly, A_{ij} is the area bounded by two circles, one of radius j and one of radius $j-1$, and further bounded in a direction perpendicular to, and coplanar with, the viewing direction by zonal boundaries at lateral displacements i and $i-1$ from the optical axis.

The above areas are taken as unitless and may be scaled to any desired dimension. Thus in general,

$$A_{ij} = \int_{i-1}^{i} \int_{[(j-1)^2 - x^2]^{1/2}}^{[j^2 - x^2]^{1/2}} \mathrm{d}y\mathrm{d}x, \quad i > j \tag{1}$$

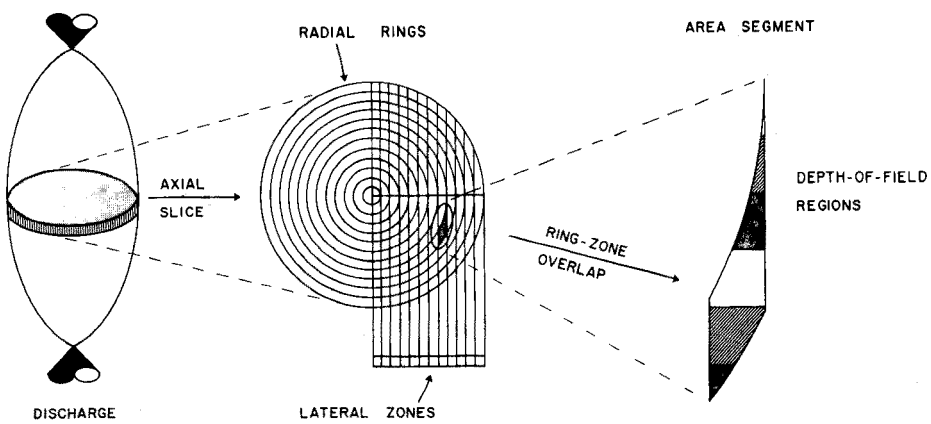


Fig. 1. Definition of terms describing spatially-resolved observation of a plasma discharge.

In the special case $i = j$, i.e. for those area segments having one edge along the diameter of the cylindrical discharge perpendicular to the optical axis, the area is

$$A_{ij} = \int_{i-1}^{i} \int_{0}^{[j^2 - x^2]^{1/2}} dy dx, \; i = j \tag{2}$$

If $i > j$, lateral zone i never overlaps ring j, so that

$$A_{ij} \equiv 0, \; i > j \tag{3}$$

Table 1 gives the general form for the above integrals. The matrix of area segments was evaluated up to $i = j = 50$. This size was arrived at by considering: (1) the resolution and field size of optical systems envisaged for use with the Abel algorithm; (2) the resolution deemed necessary to resolve spatially the salient spatial features of the spark discharges under study; and (3) to a limited extent, the core size of digital computers readily accessible for use. Consideration 3 should not override 1 or 2, as this would guarantee that information from laboratory data would be degraded in processing. In the system in use, a cylinder 5 mm in diameter can be viewed experimentally with 50-μm resolution; thus consideration of each side of the vertically symmetric discharge axis separately is precisely compatible with a 50 zone/50 ring discharge model.

A full listing of the 2500 area matrix coefficients is not included. Rather, those portions of the matrix thought to be most indicative of its properties are shown in Fig. 2 and discussed below. Both the area matrix and its inverse

TABLE 1

General form of the integrals used
(Angles in radians.)

Case A: $i > j$

$$A_{ij} = 0$$

Case B: $i = j$

$$A_{ij} = \frac{j^2}{2}\left\{\cos^{-1}\left(\frac{i-1}{j}\right) - \frac{i-1}{j}\left[1 - \left(\frac{i-1}{j}\right)^2\right]^{1/2}\right\}$$

Case C: $i < j$

$$A_{ij} = \frac{j^2}{2}\left\{\cos^{-1}\left(\frac{i-1}{j}\right) - \cos^{-1}\left(\frac{i}{j}\right) + \frac{i}{j}\left[1 - \left(\frac{i}{j}\right)^2\right]^{1/2} - \frac{i-1}{j}\left[1 - \left(\frac{i-1}{j}\right)^2\right]^{1/2}\right\}$$
$$- \frac{(j-1)^2}{2}\left\{\cos^{-1}\left(\frac{i-1}{j-1}\right) - \cos^{-1}\left(\frac{i}{j-1}\right) + \frac{i}{j-1}\left[1 - \left(\frac{i}{j-1}\right)^2\right]^{1/2}\right.$$
$$\left. - \frac{i-1}{j-1}\left[1 - \left(\frac{i-1}{j-1}\right)^2\right]^{1/2}\right\}$$

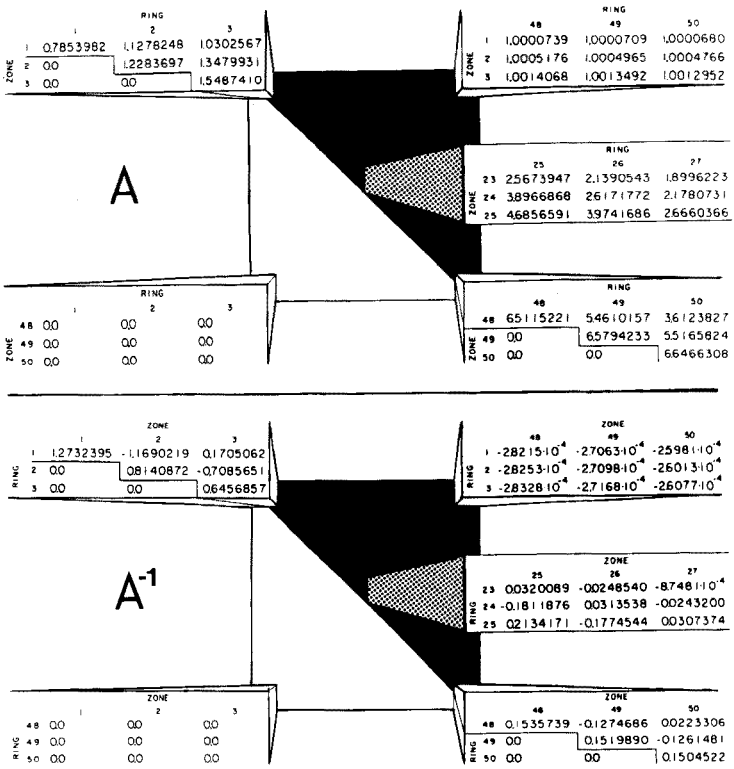


Fig. 2. Salient features of the area (A) and inverse area (A^{-1}) matrices. 45 of the 2500 elements in each matrix are shown. All elements below the diagonal of each matrix are 0.0.

are available from the authors, either as a computer-generated tabulation, or as a FORTRAN program for their creation.

The most readily evident property of the area matrix is that it is upper triangular. This is a general restatement of eqn. (3). The value of the matrix elements increases along any upper-left-to-lower-right diagonal, indicating that zone-ring overlap areas are largest in those regions of the discharge laterally farthest removed from the vertical discharge axis. In any row (along any lateral zone), the largest overlap is found for the element on the matrix diagonal. There is, however, one exception to this general observation: in lateral zone 1 ($i = 1$), the largest area value is for $j = 2$. Thus the observation of the properties of the innermost ring of the discharge is complicated not only by having to look through all the surrounding discharge layers but also by the small overlap between zone 1 and ring 1, leading to small signals related to ring 1 properties. A final property is that matrix elements in the upper right-hand corner of the matrix approach 1.0 as a limit, i.e. the overlap between zones near the optical axis and the outer rings approaches a square in shape; the apparent curvature of the discharge is largest at the core of the discharge and off axis.

Just as the area matrix may be used to compute expected intensities from predetermined radial emission distributions, so the inverse of the area matrix may be used to compute radial emission profiles from lateral intensity information. The inverse matrix is most conveniently computed at the same time as the area matrix, with both matrices being made available for later computations. Although the method for computing the matrices is described in detail, this computation need be done only once; the same matrices can be used for any discharge system of any dimension and with any number of rings and zones up to the rank of the matrix. Thus the matrices for use with a ten ring—ten zone model would be the region of intersection of the first ten rows and first ten columns in the upper left corners of the 50 ring—zone model matrices.

Properties of the inverse matrix elements A_{ji}^{-1}, (Fig. 2) include the following. First, as was the case with the area matrix, the inverse matrix is upper triangular. Secondly, along any upper left to lower right diagonal, the absolute values of the numbers decrease; thus in processing data by the inverse matrix, information gathered in nearby zones farther from the discharge axis than the zone under consideration has a pronounced effect on any given zone computations, with the strongest interaction occurring near the discharge axis. Thirdly, aside from elements A_{jj}^{-1} and A_{jj+2}^{-1}, all inverse matrix elements are negative. The physical effect of the inverse matrix is to "peel away" the light emitted in outer discharge layers to obtain information on the inner discharge structure; thus subtraction of outer ring phenomena is the predominant operation carried out by A^{-1}, which is the reason that most off-diagonal elements are negative. Fourthly, although matrix elements get smaller towards the upper right-hand corner, they never become zero. There is always a small but finite contribution of an outer ring to the intensity observed in an inner zone.

For all calculations leading to the area and inverse area matrices, double-precision calculations were performed with single-precision output available for use with processors without double precision (e.g. BASIC, in which the actual programs were written). If single-precision calculations are used, rounding-off errors of several tenths percent are observed in the inverse matrix. So that processing may be done confidently without inherent errors from the matrices, double-precision matrix calculation followed by output of as many digits as the applications program can handle is recommended. For the BASIC processor employed, 8 significant figures, as shown for most matrix elements in Fig. 2, were output, even though the confirmed precision of the matrix calculations was 14 digits. When the BASIC processor was used to multiply the area matrix by the inverse matrix, there were no errors larger than 1 part in 10^7. Thus, imprecision in laboratory data is expected to be sufficiently large ($\geqslant 0.1\%$) that imprecision in the area matrices is negligible.

The above computation of the area matrix presumes that the observational optical system ideally projects the three-dimensional discharge onto a two-dimensional focal plane. In fact, imaging changes can be seen as the discharge

is moved along the optical axis [14], and so it would be expected that some depth-of-field modification to the Abel Inversion procedure would be necessary. As the imaging changes are a function of the geometry of the optical system, the depth-of-field modifications can be applied most easily by combining the depth-of-field corrections with the area and inverse area coefficients, which are also geometric correction factors. For such modifications, the area matrices must be calculated once for each optical system rather than once for all systems as described above.

The correction procedure assumes that the resolution of the optical system parallel to the optical axis is the same as the resolution perpendicular to that axis. Thus, in addition to the radial ring—lateral zone coordinate system previously used, a grid of depth-of-field regions is added in the plane of the rings and lateral zones, but perpendicular to the lateral zones.

As illustrated in Fig. 3, the area matrix with depth of field correction is computed as follows: (1) the area segment is split into the various depth-of-field regions; (2) the area bounded by the zone, ring, and region under consideration is computed; (3) The regional area is multiplied by a depth-of-field weighting factor, specific to the region considered; (4) the overall area computed is the sum of the weighted regional areas. Thus:

$$A_{ij} = \sum_{k=1}^{50} A_{ijk} W_k \tag{4}$$

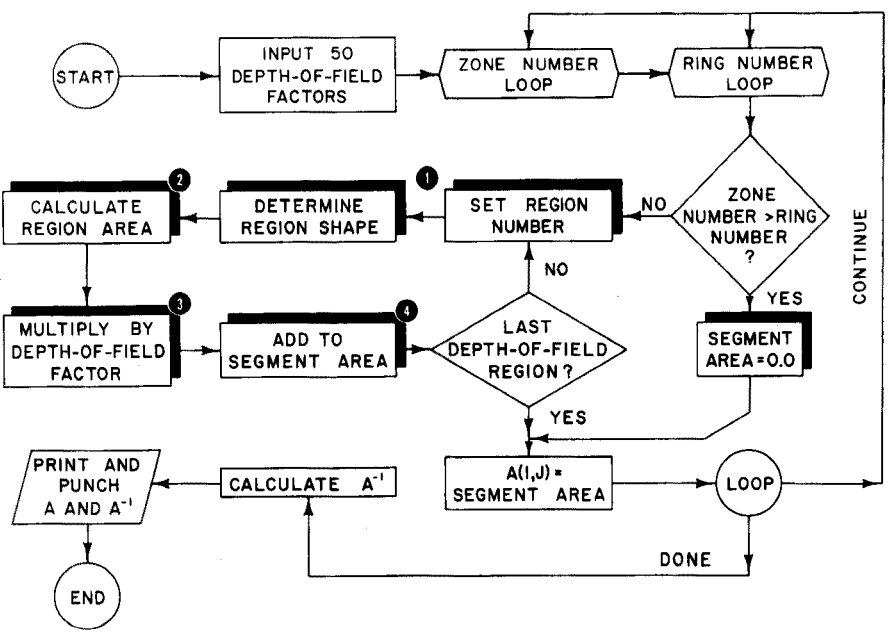


Fig. 3. Flow chart for program to calculate area matrices including depth-of-field weighting.

where W_k is the depth-of-field correction factor for region k and A_{ijk} is the area of overlap of zone i, ring j, and region k. For many regions, A_{ijk} may be 0 (e.g., in computing A_{45}, A_{45k} for $k \geqslant 5$ would be 0). Further, if W_k is set to 1.0 for all k, the area coefficients calculated without considering depth-of-field are recovered.

Table 2 and Fig. 4 show the calculation of the regional areas A_{ijk}. As before, unitless areas are computed. In Fig. 4, the area pairs (I, IV), (II, V), (III, VI), and (VII, VIII) are complementary, i.e. the area computed for I is 1 — (area computed for IV), etc. In the area segment shown in Fig. 4, the overlap types are: A_{451}, Type VI; A_{452}, Type VI; A_{453}, Type IV; A_{454}, Type I.

TABLE 2

Calculation of the regional areas A_{ijk}

Region Type, as in Fig. 4	Area of Region[a]
I	$[\frac{1}{2}(C + D) - Y(K - 1)]_{I-1}^{F}$
II	$[\frac{1}{2}(C + D) - Y(K - 1)]_{E}^{F} + E - I + 1$
III	$[\frac{1}{2}(C + D) - YG]_{I-1}^{I} + G - K + 1$
IV	$1 - [YK - \frac{1}{2}(C + D)]_{E}^{I}$
V	$[YK - \frac{1}{2}(C' + D')]_{E'}^{F'} + 1 - F'$
VI	$[YH' - \frac{1}{2}(C' + D')]_{I-1}^{I} + K - H'$
VII	$1 - [\frac{1}{2}(C' + D') - Y(K - 1)]_{I-1}^{F'}$
VIII	$[YK - \frac{1}{2}(C' + D')]_{E'}^{I}$
IX	$1 - [YK - \frac{1}{2}(C + D)]_{E}^{I} - [\frac{1}{2}(C' + D') - Y(K - 1)]_{I-1}^{F'}$

Definitions are [15]: Y = integration variable with range parallel to the optical axis; final variable for integral limits insertion. $[f(Y)]_{Q}^{R} = f(R) - f(Q)$. I = lateral zone number. J = radial ring number. K = depth-of-field region number.

$B = [J^2 - Y^2]^{1/2}$	$B' = [(J-1)^2 - Y^2]^{1/2}$
$C = Y\,[J^2 - Y^2]^{1/2}$	$C' = Y\,[(J-1)^2 - Y^2]^{1/2}$
$D = J^2 \sin^{-1}(Y/J)$	$D' = (J-1)^2 \sin^{-1}(Y/(J-1))$
$E = [J^2 - K^2]^{1/2}$	$E' = [(J-1)^2 - K^2]^{1/2}$
$F = [J^2 - (K-1)^2]^{1/2}$	$F' = [(J-1)^2 - (K-1)^2]^{1/2}$
$G = [J^2 - I^2]^{1/2}$	$G' = [(J-1)^2 - I^2]^{1/2}$
$H = [J^2 - (I-1)^2]^{1/2}$	$H' = [(J-1)^2 - (I-1)^2]^{1/2}$

[a] B, B', G', and H arise in deriving these formulae but do not appear in them.

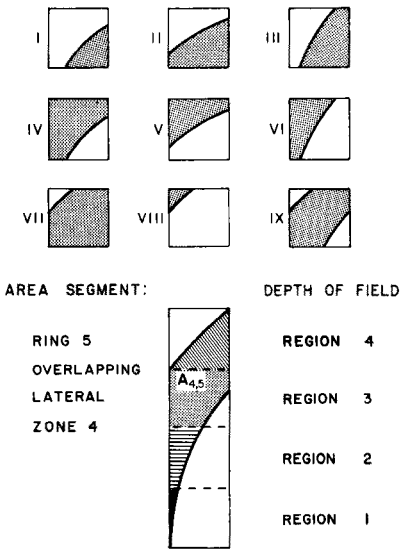


Fig. 4. Shapes possible for overlap of radial rings, lateral zones, and depth-of-field regions, together with an example of the construction of one area matrix element from several overlap regions.

The problem of which depth-of-field weighting factors to use remains. Sacks [15] and Goldstein [14] determined both experimentally and theoretically that for their optical systems (f/27 and f/17.5, respectively) weighting factors of 1.0 were indeed correct to a first order of approximation. However, neither found this conclusion totally satisfactory, and the weighting factor framework has been retained pending further insights. A promising approach [16] concluded that for radial profiles which were either uniform or peaked near the discharge axis, errors from ignoring depth-of-field effects were less than 0.25%. It remains to be seen if off-axis profiles are distorted by assuming that all depth-of-field weighting factors equal 1.0.

APPLICATION OF MATRICES TO CONCENTRIC RING DISCHARGE MODEL

For an "optically thin", i.e. purely emitting, source, the procedures for interchanging radial and lateral information are well established. The physics involved have been discussed [3, 10, 12, 16, 17]. By means of the matrices calculated above and the terms defined in Table 3, emission and intensity are related as follows

$$I_i = 2 \sum_{j=1}^{j_{emax}} A_{ij} J_i \qquad (5)$$

TABLE 3

Definition of terms

A_{ij}	Area of overlap of zone i with ring j.
A_{ji}^{-1}	Inverse area matrix element applicable to ring j, zone i.
L_{ij}	Mean absorption pathlength in zone i through ring j.
W_i	Zone width.
M_{ji}	Elements of the matrix inverse to L_{ij}.
I_i	Observed or observable intensity in zone i.
I_i'	Intensity calculated to be observed were self-absorption not present. Non-observable quantity.
I_{io}	Backlight intensity for absorption experiment in zone i.
J_j	Emission in ring j.
K_j	Absorption in ring j.
i_{max}, j_{max}	Maximum ring or zone index ($\leqslant 50$ for the matrices available).
i_{emax}, j_{emax}	Maximum ring or zone index applicable to emission portion of discharge.
I_{ib}	Background intensity in zone i.
I_{ic}	Intensity corrected for background.

$$J_j = \tfrac{1}{2} \sum_{i=1}^{i_{emax}} A_{ji}^{-1} I_i \tag{6}$$

These formulae may be modified slightly if correction for continuum background is desired. If I_i is the net observed intensity in a given zone and I_{ib} is the intensity in the same zone at an adjacent wavelength, then

$$I_{ic} = I_i - I_{ib} \tag{7}$$

where I_{ic} is the background-corrected intensity. Then, since the Abel Inversion is linear when only emission occurs, either

$$J_j = \tfrac{1}{2} \sum_{i=1}^{i_{emax}} A_{ji}^{-1} I_{ic} \tag{8}$$

or

$$J_j = \tfrac{1}{2} \sum_{i=1}^{i_{emax}} A_{ji}^{-1} I_i - \tfrac{1}{2} \sum_{i=1}^{i_{emax}} A_{ji}^{-1} I_{ib} \tag{9}$$

The first term on the right-hand side of eqn. (9) performs a lateral-to-radial conversion on the total radiation observed at the wavelength of interest; the second term inverts the background radiation. When only emission occurs, background may be subtracted either before or after inversion with numerically identical results. Also, radial profiles may be modeled or computed for the continuum radiation.

As reported previously [13], an approach very similar to that used for a strictly emitting source may be employed for sources subject to self-absorption by a spectroscopically cool cylinder surrounding the spectroscopically hot emitting core. The absorption pathlength through ring j as viewed in zone i, L_{ij}, may be computed as

$$L_{ij} = A_{ij}/W_i \tag{10}$$

where W_i is the zone width. If A_{ij} is unitless, then $L_{ij} = A_{ij}$. Similarly, if M_{ji} is the inverse of the L_{ij} matrix

$$M_{ji} = W_i A_{ji}^{-1} \tag{11}$$

If only emission or only absorption takes place in a given radial ring, then there will be a ring, j_{emax}, and a zone, i_{emax}, such that all information contained in rings or zones numbered higher than these indices will contain information on absorption only. For these zones and rings, the inversion may be described as

$$\log_{10}\frac{I_{i0}}{I_i} = 2\sum_{j=j_{emax}+1}^{j_{max}} L_{ij}K_j \tag{12}$$

$$K_j = \tfrac{1}{2}\sum_{i=i_{emax}+1}^{i_{max}} M_{ji}\log_{10}\frac{I_{i0}}{I_i} \tag{13}$$

where I_{i0} is the intensity of a backlighting source viewed through zone i in the absence of self-absorption.

For $i \leqslant i_{emax}$ and $j \leqslant j_{emax}$, the emitting portion of the discharge is viewed through the absorbing shell. Thus the discharge acts as a light source viewed through an absorption cell. Overall, the observed intensity (with no external backlight employed) can be described as

$$I_i = \left(2\sum_{j=1}^{j_{emax}} A_{ij}J_j\right)\cdot 10^{-\left(\sum_{j=j_{emax}+1}^{j_{emax}} L_{ij}K_j\right)} \tag{14}$$

and conversely the emission in any ring may be calculated by

$$J_j = \tfrac{1}{2}\sum_{i=1}^{i_{emax}} A_{ji}^{-1}I_i\cdot 10^{\sum_{j=j_{emax}+1}^{j_{max}} L_{ij}K_j} \tag{15}$$

For modeling and data analysis, it is often desirable to be able to examine intermediate steps in the calculation as well as the final results. For determining the emission in the emitting rings, eqn. (15) does not allow intermediate inspection, but goes directly from observation to ring emission.

Equation (15) may be broken into two steps to clarify the calculation and permit intermediate examination of the intensity data after compensation for self-absorption. If I_i' is the intensity which would be observed in zone i in the absence of self-absorption, then

$$I_i' = I_i \cdot 10^{\sum_{j=j_{emax}+1}^{j_{max}} L_{ij} K_j} \tag{16}$$

and eqn. (15) may be rewritten

$$J_j = \tfrac{1}{2} \sum_{i=1}^{i_{emax}} A_{ji}^{-1} I_i' \tag{17}$$

which is identical to eqn. (6) if all the K_j values are zero, i.e. in the absence of self-absorption.

For designing algorithms to allow facile conversion from radial profiles to lateral (observable) profiles and vice versa, the equations found to be most comprehensive and yet to allow full study of intermediate calculation steps are: 5, 12, 13, 14, 16, and 17. The flow-chart in Fig. 5 shows how these equations were grouped in the modeling software employed earlier [13]. Figure 5A is for conversion of radial profiles to lateral profiles, and Figure 5B is for lateral-to-radial profile conversion; the multiple entry provides flexible computation and interaction. The equations are arranged one to a block. Block 1 corresponds to eqn. (5), block 2 to eqn. (12), and block 3 to eqn. (14). Similarly, block 5 corresponds to eqn. (13), block 6 to eqn. (16), and block 7 to eqn. (17) or, when no absorption is present, to eqn. (6). Because of the

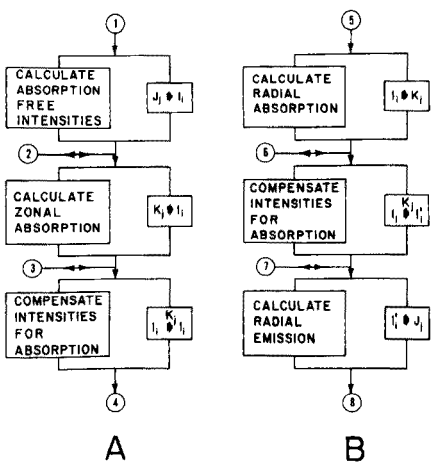


Fig. 5. Multiple entry routines for interconversion of radial and lateral spatially resolved information. Routine A converts from radial profiles to lateral profiles, whereas routine B converts from lateral profiles to radial profiles.

multiple-entry structure, other data manipulations such as background subtraction or data smoothing may be performed at any desired step in the procedure, and, with appropriate choice of language and control statements, in real time. It is convenient to write all the modeling software in BASIC.

To avoid distortion of the inverted profiles, background correction for laboratory data must be done after the effects of self-absorption have been accounted for. Thus in Fig. 5B, background subtraction may occur either at entry point 7 (subtract background intensities) or at exit point 8 (subtract background radial emission). As shown in eqns. (7—9), the two approaches give identical results. Attempting to compensate for background before entry point 7 leads to difficulties [13].

There may be physical situations where it is necessary to consider emission and absorption occurring in exactly the same position. Algorithms for handling simultaneous emission and absorption in a conceptual framework similar to the above are presently being designed, in conjunction with laboratory application of the models presented.

MODELING OF OPTICAL MISALIGNMENT AND DISCHARGE WANDER

One of the major experimental difficulties in obtaining spatially resolved spectra is the stabilization and alignment of the discharge with the optical system so that the volume viewed is reproducible and at a known location within the discharge. Discharge wander may be regarded as an irreproducible optical misalignment in which each discharge passes through a unique spatial path, and the observed light intensity is spatially averaged as a result. Thus modeling of wander may be performed by an algorithm consisting of three parts: (1) an algorithm which operates on lateral data as it would be viewed in a properly aligned, spatially stable system to produce data laterally displaced from proper alignment, i.e. shifted with respect to the instrument optical axis; (2) a routine to vary the amount by which the data is shifted so that various distribution patterns for the wander may be modeled; and (3) a normalization routine to give a properly scaled value in each zone for the time-integrated observed intensity. This allows comparison on the same scale as the number of discharges used in modeling changes.

The routines described here will work only for purely emitting species distributions. Reasons for difficulties in modeling self-absorbed systems will be discussed below. Material illustrating the algorithm is shown in Fig. 6.

In view of the assumed cylindrical (or at least bilateral) symmetry of the discharge, observation of the discharge on only one side of the discharge/optical axis should suffice to elucidate fully the discharge structure. Thus for the radial profile shown in Fig. 6(a) it is sufficient in a stable discharge to look only at that portion of the lateral profile shown in Fig. 6(b), where the vertical axis in the inset corresponds to both the discharge and instrument optical axes. In such a case it is assumed that the mirror image of the lateral profile (b) would appear on the other side of the axis. Thus when wander

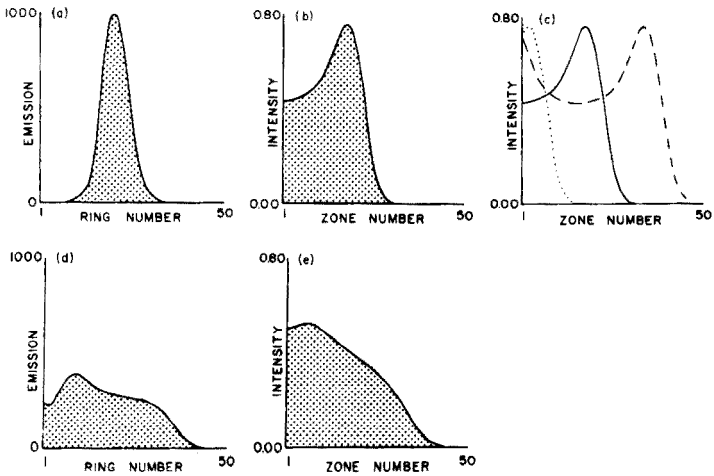


Fig. 6. Computational steps in modeling discharge-wander for a purely emitting source. (a) Radial information showing a true radial profile. (b) Lateral information showing a wander-free lateral profile. (c) Individual laterally displaced profiles; displacements are +15 zones (---), ±0 zones (—), and —15 zones (····). (e) Normalized lateral data with ±15 zone wander. (d) Calculated radial profile.

or misalignment occur, either portions of the lateral profile observed in a stable aligned discharge are obscured as they pass to the left of the axis, or portions of the discharge normally not viewed appear in the viewing field so that the apparent diameter of the discharge increases.

Figure 6(c) shows three lateral profiles which would result from various amounts of wander in a single discharge. If the lateral profile has to be shifted by S zones, then the intensity in zone N after the shift is completed may be determined from the unshifted data as follows:

$$I_S(N) = I(N-S) \qquad S \leqslant 0 \text{ and } N-S \leqslant i_{\max} \tag{18}$$

or

$$S \geqslant 0 \text{ and } N-S \geqslant 1$$

$$I_S(N) = 0 \qquad S \leqslant 0 \text{ and } N-S > i_{\max} \tag{19}$$

$$I_S(N) = I(S-N+1) \qquad S \geqslant 0 \text{ and } N-S < 1 \tag{20}$$

In all cases, S and N are integers and N may range from 1 to i_{emax}; I and I_S are, respectively, the intensities before and after shifting in the zone indicated within the parentheses. These formulae account not only for shifting and for the symmetry axis but also assume for a negative shift (e.g. Fig. 6(c)) that there is no intensity beyond zone i_{emax}.

If the wander distribution is such that the discharge axis for any random discharge is uniformly distributed ±15 zones from the optical axis, the total intensity distribution is as shown in Fig. 6(e). The uniform distribution

was obtained by scaling and offsetting the pseudo-random numbers calculated by the RND function in Univac BASIC. Figure 6(d) shows the radial profile obtained by inverting the lateral profile in Fig. 6(d). Inversion was performed by using entry point 7 as shown in Fig. 5. The qualitative differences between (a) and (d) in Fig. 6 are readily evident. Thus manipulation of lateral data by appropriate shifting of subscripts clearly shows the effects of unstable discharges on the inversion of lateral information to radial information.

Many distributions besides the uniform pseudo-random one shown here may be used. If the lateral distribution is merely offset by S zones, the effect of gross optical misalignment may be modeled. If the shift is applied to a radial profile rather than a lateral profile, the effects of discharge expansion and contraction may be modeled. A Gaussian random distribution may be obtained by using appropriate software as well.

These algorithms in combination contain several limitations. First, there is no provision for wander in a direction parallel to the optical axis in addition to the situation modeled, i.e. wander perpendicular to the axis. This stems from the uniform processing of data regardless of relative position along the optical axis, a circumstance derived from the use of all depth-of-field weighting factors equal to 1.0 in the area matrix calculation. In a related limitation, there is no provision for modeling a tilt of the discharge with respect to the line formed by instrumental slits. This non-ideality could be modeled if the area matrix were calculated for an ellipsoidal discharge rather than a cylindrical one (a plane slicing through a cylinder traces out an ellipse in general; a cylindrical slice is obtained only when the plane is perpendicular to the cylinder axis).

Self-absorption effects can be modeled only to a limited extent. If self-absorption is the only process occurring, shifting as described above could be used with little modification to model the effects of wander. However, if both emission and absorption occur, the modeling procedure breaks down. No longer is there a defined boundary to the emission or an inner spatial limit to the absorption. This violates the assumption of spatial segregation of emission and absorption required by the equations given previously. By implication, even ± 1 zone wander is sufficient to preclude valid use of the spatially segregated emission—absorption model for the discharge if self-absorption is present. Analysis of emission and absorption in an unstable discharge demands a model which allows both emission and absorption in the same spatial region.

Finally, there are the implicit assumptions that no chemical changes occur which would change the radial profiles when wander occurs, and that wander is a discharge property which has no effect on any other property and is not influenced by any other characteristic. In fact, observable changes in discharge properties occur in conjunction with changes in the degree of discharge wander [19, 20]. Distortions in radial and lateral profiles as a result of wander are therefore expected to be greater in experimental work than is demonstrated by the present modeling procedure.

The financial support of the National Science Foundation for a traineeship for AS and for computing, and support from the Graduate School and Department of Chemistry, University of Wisconsin, are acknowledged. The assistance of Al Christoph, Charles Green, Chuck Hutchins, and Dean Stueland with the computer software, and of Pat Brinkman in the preparation of the figures is appreciated. This study is partly based on work completed by R. D. Sacks and S. A. Goldstein in partial fulfilment of the requirements for Ph.D. (Chemistry).

REFERENCES

1 W. S. Eaton, Ph.D. Thesis, University of Wisconsin, 1975 (University Microfilms 75-12893).
2 J. P. Walters and S. A. Goldstein, How and What to Sample in the Analytical Gap, Sampling, Standards, and Homogeneity, ASTM 5TP 540, (1973) 45.
3 P. W. J. M. Boumans, Theory of Spectrochemical Excitation, Hilger and Watts, London, 1966.
4 C. D. Maldonado and H. N. Olsen, J. Opt. Soc. Am., 56 (1966) 1305.
5 H. W. Emmons, Phys. Fluids, 10 (1967) 1125.
6. L. I. Grechikhin and V. D. Shimanovich, Opt. Spectrosc., 15 (1961) 358.
7 J. F. Bott, J. Quant. Spectrosc. Radiat. Transfer, 6 (1966) 807.
8 R. J. Decker and P. A. McFadden, Spectrochim. Acta, Part B, 30 (1975) 1.
9 D. J. Kalnicky, R. N. Kniseley, and V. A. Fassel, Spectrochim. Acta, Part B, 30 (1975) 511.
10 D. J. Kalnicky, V. A. Fassel, and R. N. Kniseley, Appl. Spectrosc., 31 (1977) 137.
11 R. D. Sacks and J. P. Walters, Anal. Chem., 42 (1970) 61.
12 S. A. Goldstein and J. P. Walters, in preparation.
13 A. Scheeline and J. P. Walters, Anal. Chem., 48 (1976) 1519.
14 S. A. Goldstein, Ph.D. Thesis, University of Wisconsin, 1973 (University Microfilms 73-20996).
15 R. D. Sacks, Ph.D. Theiss, University of Wisconsin, 1969 (University Microfilms 70-3687).
16 K. C. Lapworth, J. Quant. Spectrosc. Radiat. Transfer, 16 (1976) 357.
17 W. J. Pearce in H. Fischer and L. C. Mansur (eds.), Conference on Extremely High Temperatures, Wiley, New York, 1958, p. 123.
18 W. J. Pearce in P. J. Dickerman (ed.), Optical Spectrometric Measurements of High Temperatures, University of Chicago Press, Chicago, 1961, p. 125.
19 D. M. Coleman, Ph.D. Thesis, University of Wisconsin, 1976 (University Microfilms, 77-3393).
20 D. M. Coleman, J. P. Walters, J. Appl. Phys. Lett., in press.

Analytical Pyrolysis

Proceedings of the Third International Symposium held in Amsterdam, September 7 - 9, 1976

C. E. ROLAND JONES and CARL A. CRAMERS *(Editors)*

This symposium is particularly noteworthy because of the emphasis given to the newly emergent technique of pyrolysis/mass spectrometry. The large number of papers devoted to this technique at the meeting are an indication of the impetus which this recent development has given to analytical pyrolysis.

These Proceedings provide examples of a diversity of applications of pyrolysis/gas chromatography and pyrolysis/mass spectrometry ranging from geochemical exploration through energy resource studies to the elucidation of biopolymers and complex synthetic resins. The thirty-four papers give perspective to the current state of the fields, as well as reporting on the most recent developments in them. The introductory contributions in the sessions, provided by prominent figures in the particular fields, summarize the position to date before revealing the latest trends in the authors' own work. It could be said that each session was a miniature symposium in itself.

1977 x + 424 pages US $39.25/Dfl. 96.00 ISBN 0-444-41558-0

The Dutch guilder price is definitive. The US $ price is subject to exchange rate fluctuations.

ELSEVIER

7014

Quadrupole Mass Spectrometry and its Applications

edited by **PETER H. DAWSON,** National Research Council of Canada.

1976. xxii + 350 pages. US $49.75/Dfl. 129.00. ISBN 0-444-41345-6

This is the first comprehensive account of quadrupole mass spectrometry. While its many contributors provide a broader-than-usual viewpoint, it is, nevertheless, a systematic text. It begins with simple qualitative descriptions of the mass filter, the monopole, the quadrupole ion trap and related time-of-flight spectrometers. It proceeds to an exploration of their particular advantages, disadvantages and applications. Experimental design and performance is discussed in detail. The theoretical treatment includes computational design techniques such as the recently developed utilisation of phase-space dynamics. Although there have been countless routine applications of quadrupole mass spectrometry, there are an unusual number of individualized applications in both science and technology which require specially designed or modified instruments. This book will therefore be of interest and value to many users for whom a knowledge of quadrupole design, performance and limitations may be essential.

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

P.O. Box 211, Amsterdam, The Netherlands

Distributor in the U.S.A. and Canada:
ELSEVIER/NORTH-HOLLAND, INC.,
52 Vanderbilt Ave., New York, N.Y. 10017

The Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

VOL. 95, No. 1/2 ANALYTICA CHIMICA ACTA September 15, 1977
(Computer Techniques and Optimization, Vol. 1, No. 1/2)

CONTENTS