Vol. 103 No. 4 December 15, 1978
(Computer Techniques and Optimization, Vol. 2 No. 4)
Int. Conf. Computers and Optimization in Analytical Chemistry, Amsterdam, April 5–7, 1978
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 1978 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	96/1	96/2	97/1	97/2	98/1	98/2	99/1	99/2	100	101/1	101/2	102
Section on Computer Techniques and Optimization			103/1			103/2			103/3			103/4

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. The section deals with the following topics: Computerized acquisition, processing and evaluation of data. Computerized methods for the interpretation of analytical data including chemometrics, cluster analysis, and pattern recognition. Storage and retrieval systems. Optimization procedures and their application. Automated analysis for industrial processes and quality control. Organizational problems.

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, Universität Bern, Pharmazeutisches Institut, Sahlistrasse 10, CH-3012 Bern, 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. 102, p. 253 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.

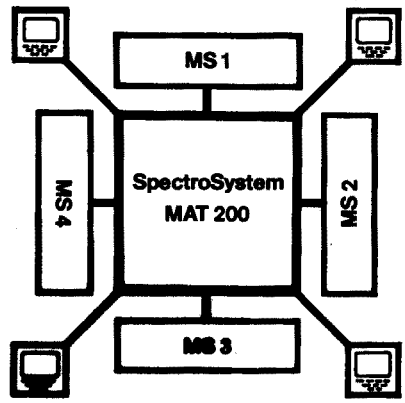
Subscriptions. Subscriptions should be sent to: Elsevier Scientific Publishing Company, P.O. Box 211, 1000 AE Amsterdam, The Netherlands. The section on *Computer Techniques and Optimization* can be subscribed to separately.

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 8 volumes in 1978. The subscription for 1978 (Vols. 96–103) is Dfl. 1000.00 plus Dfl. 120.00 (postage) (Total approx. US $486.96). The subscription for the *Computer Techniques and Optimization* sections only (Vol. 103) is Dfl. 125 plus Dfl. 15.00 (postage) (Total approx. US $60.87). 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. 1–95) except Vols. 23 and 28 are available at Dfl. 144.00 (U.S. $63.00), plus Dfl. 10.00 (U.S. $4.35) postage and handling, per volume.

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

Customers in the U.S.A. and Canada who wish to obtain additional bibliographic information on this and other Elsevier journals should contact Elsevier/North Holland Inc. Journal Information Center, 52, Vanderbilt Avenue, New York, NY 10017. Tel: (212) 867-9040.

Varian MAT's most powerful mass spec data system



The SpectroSystem MAT 200 is the only data system that offers both multi user operation and simultaneous multi instrument operation

Four users simultaneously utilize one computer which:

- allows interactive and/or automatic analysis processing
- performs intelligent analysis clean up
- does library search and comparison
- supports program development in FORTRAN
- uses a well accepted general purpose operating system

Four mass spectrometers are simultaneously on-line with one computer which:

- acquires their data through intelligent interfaces
- stores data in analysis files
- generates real-time reports (graphic and numeric) keeping the operator informed about the progress of the experiment
- supports unattended operation by way of preprogrammed tasks (batch operation)
- contains modern 16 bit technology

The SpectroSystem MAT 200 is very flexible and uses advanced computer hardware and software (DEC computer PDP-11/34 with floating point processor, operating system RSX-11 M). The application software is written in FORTRAN IV PLUS.

Varian MAT solves your analytical problems. Talk to our experts – ask for more information.

Varian MAT – the leading name in mass spectrometry

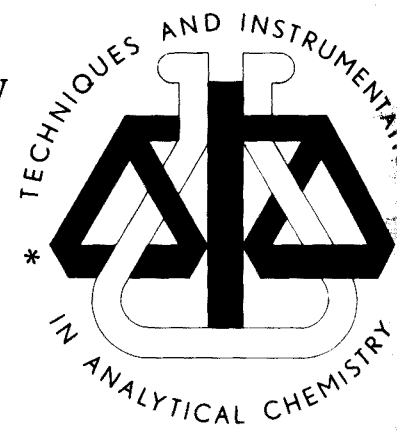
Varian MAT GmbH
Postfach 14 40 62
D-2800 Bremen 14
F. R. Germany

varian

Varian MAT
Mass Spectrometry
25 Hanover Road
Florham Park, N. J. 07932
USA

200e178

Evaluation and Optimization of Laboratory Methods and Analytical Procedures



A Survey of Statistical and Mathematical Techniques

D.L. MASSART, A. DIJKSTRA *and* L. KAUFMAN.

with contributions by S. Wold, B. Vandeginste *and* Y. Michotte

Techniques and Instrumentation in Analytical Chemistry - Volume 1

This book provides detailed treatment, in a single volume, of formal methods optimization in analytical chemistry. It is a comprehensive and practical handb which no analytical laboratory will want to be without.

All aspects of optimization are discussed, from the simple evaluation of procedu to the organization of laboratories or the selection of optimal complex analyti programmes. Quantitative discrete analysis as well as qualitative and continu measurement techniques are evaluated.

The book consists of 30 chapters divided into 5 main parts.The main sections Evaluation of the Performance of Analytical Procedures, Experimental Optimizati Combinatorial Problems, Requirements for Analytical Procedures, and Syste Approach in Analytical Chemistry.

This work will be of practical value not only to those involved with optimiza problems in analytical chemistry, but also to those in related fields such clinical chemistry or specialized fields such as chromatography. Becaus discusses the application of many mathematical techniques in analytical chemi this book will also serve as a general introduction to the new field of Chemomet

Oct. 1978 xvi + 596 pages US $57.75/Dfl. 130.00 ISBN 0-444-41743-5

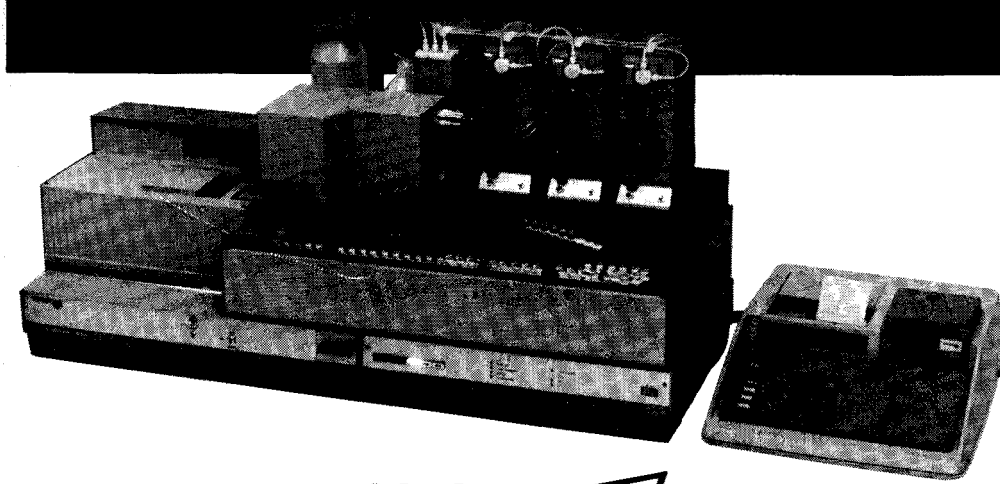
ELSEVIER

P.O. Box 211,
1000 AE 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.

It's Child's Play



with the new Vitatron Programmable Analyser PA 800

This is the easy way to run your **enzyme activities substrates** (kinetic and end-point methods **emit's®**

1 insert program card
2 set dispensers
3 set wavelength
4 load reagents
5 push the startbutton

Rinsing, calibrations and calculations are performed automatically. The PA 800 with its programmable calculator is a really versatile instrument offering a variety of applications with a choise of manufacturer's reagent.

Other fine features like the low running costs, reliability, accuracy and simple handling, makes the PA 800 analyser 'second to none'.

Vitatron Scientific B.V.

P.O. Box 100
6950 AC DIEREN
The Netherlands
Dealers in 60 countries

Medica '78
22-25 November
Stand nr. 5050

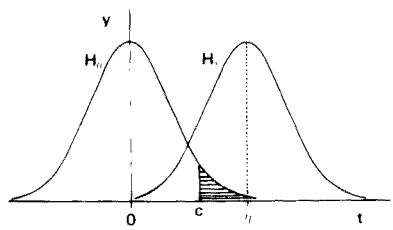
®Emit is a registered trade mark of Syva, Palo Alto, California, U.S.A.

Statistical Treatment of Experimental Data

by 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.*

Physical Sciences Data, Vol. 2

First published in 1977 and now reprinted with some minor revisions, 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.

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.

CONTENTS: Chapters: 1. Introduction. 2. Probability. 3. Random Variables and Sampling Distributions. 4. Some Important Probability Distributions. 5. Estimation. 6. Confidence Intervals. 7. Hypothesis Testing. 8. Tests on Means. 9. Tests on Variances. 10. Goodness of Fit Tests. 11. Correlation. 12. The Straight Line Through the Origin or Through Some Other Fixed Point. 13. The Polynomial Through the Origin or Through Some Other Fixed Point. 14. The General Straight Line. 15. The General Polynomial. 16. A Brief Look at Multiple Regression. Appendices: 1. Drawing a Random Sample Using a Table of Random Numbers. 2. Orthogonal Polynomials in x. References. Index.

1977 1st revised reprint 1978 xiv + 382 pages US $39.25/Dfl. 90.00
ISBN 0-444-41725-7

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.

7046A

SPECIAL ISSUE

COMPUTERS AND OPTIMIZATION IN ANALYTICAL CHEMISTRY

Proceedings of a Symposium held in Amsterdam, April 5–7, 1978

Analytica Chimica Acta, 103 (1978) 271
Computer Techniques and Optimization

FOREWORD

Optimization is a keyword in analytical chemistry. The optimization of methods, procedures, strategies and laboratories has always been a principal objective of analytical chemists. The computer has recently added a new dimension, enabling rationalization and automation of a good deal of routine work. But the power of the computer is not limited to mechanization or automation. The application of mathematics, statistics and other chemometric techniques, formerly too time-consuming, has now come within reach for routine purposes. Interest in data retrieval and library search methods is still increasing, while the application of the microprocessor is completely changing the image of analytical instrumentation.

The International Conference on Computers and Optimization in Analytical Chemistry, held in Amsterdam in April, 1978, was organized — under the auspices of the Royal Netherlands Chemical Society, Analytical Division — to present the state of the art in these fields and to stimulate new developments by bringing together scientists interested in chemometrics, automation, computers, data retrieval and other related subjects. An international scientific committee, led by A. Dijkstra, G. Kateman and H. C. Smit assembled the scientific program, and many excellent papers were presented. Most of these papers are contained in this special issue of Analytica Chimica Acta/ Computer Techniques and Optimization.

The conference certainly fulfilled its objectives and it is hoped that this publication of the major part of the proceedings will act as a stimulant for further interesting developments.

H. C. Smit
Laboratory for Analytical Chemistry,
University of Amsterdam.

Analytica Chimica Acta, 103 (1978) 273—288
Computer Techniques and Optimization

PROBLEMS IN DATA RETRIEVAL SYSTEMS FOR ANALYTICAL SPECTROSCOPY

JURE ZUPAN
Institute of Chemistry Boris Kidrič, Ljubljana (Yugoslavia)
(Received 3rd May 1978)

SUMMARY

Some of the existing retrieval systems for analytical purposes based on different spectrometric data are described and compared with systems under development. Requirements concerning input data, coding of fragments or complete chemical structures, preprocessing of input data, and evaluation of the retrieved results are discussed, and some examples are given. Some trends for future developments in this field are outlined.

In the last four decades, two major events have drastically affected analytical chemistry: first, the advent of commercially available automated instrumentation for various purposes, and secondly the intensive use of large, mini and micro computers for handling chemical information of any type [1]. The effects have been particularly dramatic for analytical procedures based on various types of molecular spectra. In fact, the use of computers was made almost inevitable by the enormous amount of spectral data. And, in feedback of fast data or signal handling, only computers can manipulate the sophisticated coupled spectrometers, such as g.c.—m.s., that produce even more data than the simple instruments. The fast growing computer capability for data handling focused attention on the construction of powerful information systems which could carry out the process of structure elucidation based on spectral data. Chemical information systems existed long before computers penetrated into chemistry but their power was greatly limited by the ingenuity of the user: the correlations between structures and spectral features are so precarious that they could be utilized in full only by specialists.

The development of the computer-assisted information systems was rather straightforward. First, some small and very specialized computer- and spectrometer-oriented systems were developed, but more general fully computerized sophisticated systems were later constructed or are still under development. To date, more than a hundred specialized retrieval systems for spectrometric data have been developed. Collections of references to the most important ones are available in the recent literature [2—5]. Whilst the number of simple or complex retrieval systems dealing with only one type of spectrometry is very large, the number of complex systems that can handle data simultaneously

from different spectroscopic methods in order to define the structure of unknown compounds is still less than ten [2, 3, 5—8].

The main task of this paper is to indicate the general ideas and the problems that must be, or have already been, solved in order to achieve the final goal of interpreting the structures of unknown compounds. The similarities or differences in the systems available will be stressed where this seems useful.

Unfortunately, there are two points of view for many problems: the programmer's and chemist's. Often what seems very important to the first appears of minor importance to the second, and vice versa. Some problems which should be of interest to both sides will be highlighted.

DESCRIPTION OF THE SYSTEMS

Figure 1 shows a flow chart for the structure solving process based on spectral data. Such a process need not be automated at all — in fact, it represents the usual thought process of a spectroscopist interpreting an unknown

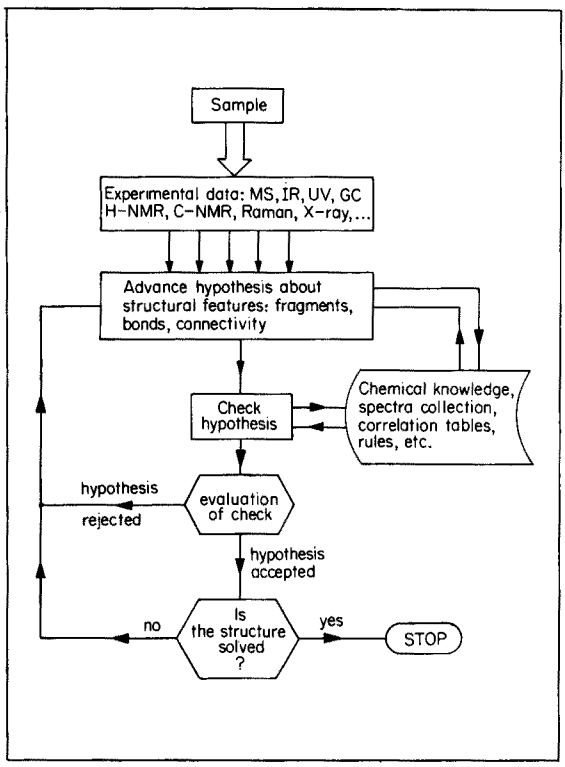


Fig. 1. Flow chart of an information system based on different spectrometric methods. From the solved structure the molecular formula, name and other properties can easily be retrieved.

structure. It is not surprising that this general concept (though not all its details), is recognizable in computerized systems — including the very complex systems [2, 3, 5—8]. Although the approaches to constructing real chemical information systems vary quite considerably from one author to another, the basic skeleton remains the same.

Some basic data about the systems in question are listed in Table 1. The first system, the NIH—EPA Chemical Information System [2] is the only complete on-line system, which makes its design philosophy (especially from the computational point of view) quite different from the rest. It is at the moment the only commercially accessible system and is used by more than 250 laboratories representing 150 different organizations in North America and Europe; it offers the largest assortment of different data files. In its present form, more than 150 000 items are stored in 40 files. Together with those specified in Table 1, there are about 30 000 x-ray data sets, and data on more than 40 000 pollutants, pesticides and other hazardous materials, 20 000 toxic substances, 9 000 items from the Merck Index, and so on. The backbone of this system is the substructure search [9—11] that makes it possible to retrieve from any of the above files all compounds and the data related to them which contain a structure or fragment pre-specified by the user's query. The advantage of the substructure search will be discussed later. The next three systems, by Clerc et al. [8, 12—14], Koptjug [7] and Zupan et al. [5], are in principle library searches but have quite different file organizations, evaluations of retrieved results, and I/O handling. The last two systems by Gray [6] and Gribov et al. [3, 15, 16] could be regarded as structural group—spectral feature correlation programs. The concept of these two programs is rather different from the previous ones, resembling in some points the DENDRAL [17—19] system. First, the extraction of structural fragments that are possible and those that are forbidden is done in the form of a "goodlist" and a "badlist", and then, in a "jig-saw" manner, a complete structure is composed that should

TABLE 1

Technical data about some complex information systems

Authors	System	Computer	M.s.	I.r.	H-n.m.r.	C-n.m.r.	U.v.	Others
			Spectra in files					
Heller et al. [2]	NIH-EPA Chemical Inf. System	DEC 10	26 000	—	Program GINA	4 000	—	130 000
Clerc et al. [4]	OCETH System Zürich	CDC 6500/6400	10 000	2 000	—	4 000	2 000	—
Koptjug [7]	CODATA Novosibirsk	MINSK 32	10 000	60 000	15 000	—	5 000	—
Zupan et al. [5]	COSMOSS Ljubljana	CDC CYBER 172	16 000	102 000	—	4 000	—	—
			Spectra—fragment correlations					
Gray [6]	Cambridge England	IBM 370/165	20	20	20	—	20	molecular formula
Gribov et al. [3]	STREC Moscow	MINSK 22	270	80	150	—	20	molecular formula

be in accordance with all the spectral features and molecular formula that were input initially.

Without question, a system that would unite the best strategies, data, and procedures applied in the systems mentioned would be far better than any single one. Thus comparison, evaluation and exchange of ideas is the best way towards an ideal chemical information system.

The term "chemical information system" covers much more than simple retrieval of stored information as spectra or structure of the desired compound. Such a system should provide the user with options for combinations of retrieved data in such a way that the ultimate goal — the complete identification of the compounds (names, structures, properties, sources, etc.) — will be accomplished with as high a success rate as possible. In a computerized chemical information system where not only must the shortest procedure be well defined and highly efficient, but must also fit properly into the entire organization scheme, a knowledge and proper choice of computational methods is of very great importance. In the following pages, general descriptions of search algorithms, evaluation of retrieved spectra, correlation of spectral features with structural fragments, coding of chemical structures, decision-making processes, mixtures, and data input/output used in chemical information systems are given.

SEARCH ALGORITHMS

The most widely used (and misused) and time-consuming routines in all information systems are without doubt the search algorithms. The choice of a proper search algorithm for a specific task will often lead to a substantial time gain, often more than 40% [20, 21]. It is necessary to keep in mind that in a complex chemical information system, data should be retrieved on the basis of different key types such as molecular formulae, spectral features, registry numbers, WLN's, or complete spectral data, and there is no single algorithm that would fit best in all these cases.

The efficiency of the search, however, depends not only on the algorithm employed but also on the file organization. For some kind of requests, especially for on-line work, inverted files [22] are generally more useful than regular sequential files (one large record per compound). However, it depends on the capability of the machine if some types of searches can be done or not. If, for example, a high-level compiler does not allow random access files, or if the machine is too small, the programs must be written in assembler, which makes programing more difficult and time-consuming.

The decision about the type of search algorithm to be implemented should be based primarily on the goal of the particular search rather than on the hardware available. Figure 2 shows some possible decisions for different types of search, but any retrieval algorithm must be tested very thoroughly before use [21]. As the most time-consuming algorithm, the sequential search should be avoided wherever possible, but in some cases it cannot be avoided. The

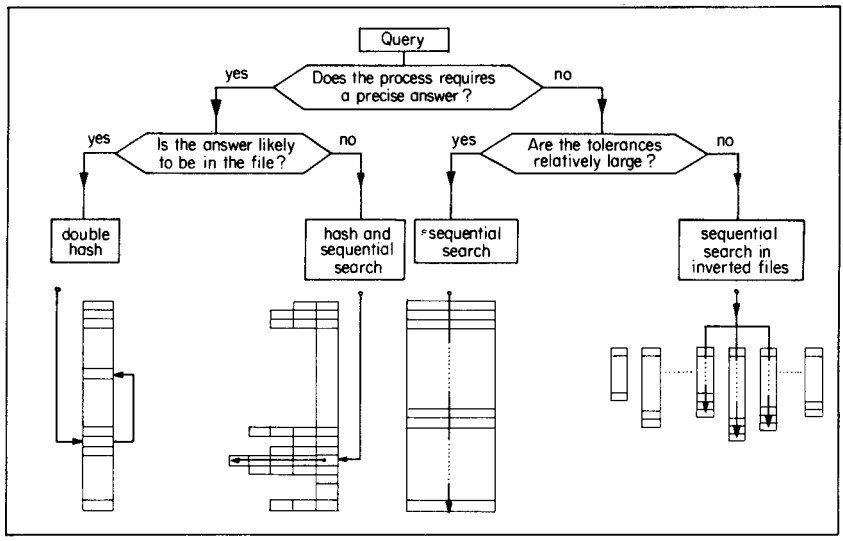


Fig. 2. Search algorithms for different types of retrieval.

fastest way of retrieving an answer is provided by hash coding algorithms [21, 23] and these should be used whenever possible. Unfortunately, the "birthday paradox" [24] enters the picture here to its greatest extent, so that the possibility of collisions caused by the same hash address from two different keys must be considered very seriously.

An excellent review of search algorithms has been published by Knuth [21].

EVALUATION OF RETRIEVED SPECTRA

A very important part of any information system is the evaluation of the results of each step (or of the whole procedure) towards the final structure determination of the compound investigated. In a "classical" library search system, the results retrieved are evaluated at the end, after the search run is complete, mainly by using algorithms with empirical constants. The evaluation is done by comparison of the spectra of the unknown compound with the retrieved spectra. The closest agreement between two spectra indicates the closest similarity in structures. Figure 3 shows how a rating algorithm [5] can yield valuable structural information even when the correct answer is not found in the search run, because of miscoding of either the input data or the data in the collection. The rating algorithm used for this particular case was combined from two ratings r_p and r_i, for positions and intensities, respectively:

$$r_{total} = A_1 r_p + A_2 r_i \qquad (1)$$

where A_1 and A_2 have the values 1, 0, 1/2 and 0, 1, 1/2 for infrared, mass, and ^{13}C-n.m.r. spectroscopy, respectively. The ratings r_p and r_i are defined as follows:

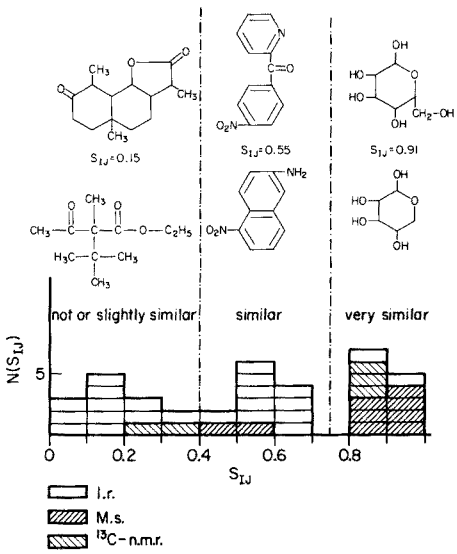


Fig. 3. The similarity between the structures of the top rated and requested compounds for unsuccessful searches. To give a visual impression of the meaning of the terms "very similar", "similar", and "slightly similar", examples are drawn with the corresponding similarity coefficients $s_{I,J}$. The similarity coefficient $s_{I,J}$ between two structures I and J was obtained by using both WLN(I) and WLN(J).

$$r_p = B_1 n_1 + B_2 n_2 + B_3 n_3 + B_4 n_4 \tag{2}$$

where B_i are empirical estimated integer constants between 1 and 100, n_1 and n_2 are, respectively, the numbers of missing and extra peaks, and n_3 and n_4 are the numbers of peaks within the tolerance intervals $\pm a$ and $\pm a$ to $\pm 2a$, respectively [25];

$$r_i = \sum_{j=1}^{k} d_j^2 \tag{3}$$

where k is the number of peaks and d_j is the difference in intensity between two corresponding peaks in the input and retrieved spectra. Lower rating means better similarity.

The actual comparison between the chemical structures (the "similarity" coefficients for structures) was evaluated by using both Wiswesser Line Notations [5, 26].

A very neat approach to the evaluation of structural similarity on the basis of spectra comparison that could be used for any kind of spectroscopy in the same manner (by changing only the weighting) was developed by Clerc et al. [12—14]. Figure 4 shows the special organization of the library file with coded structural features for each compound in the file, together with a set of weights

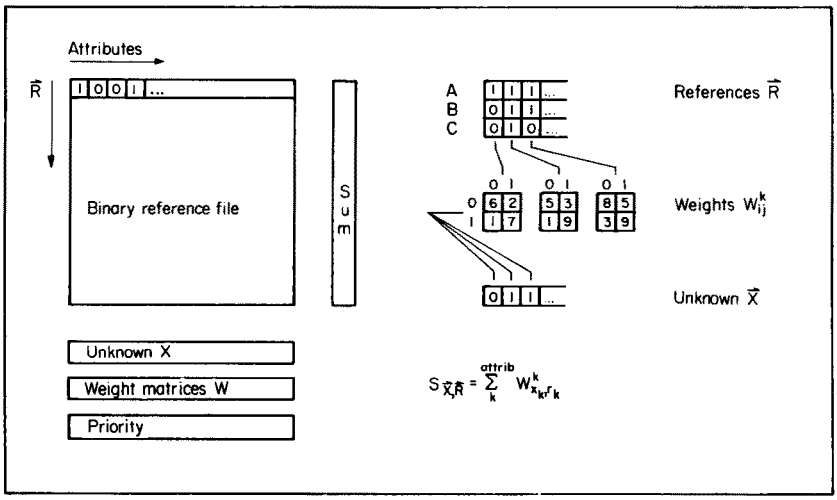


Fig. 4. Organization of the OCETH library files.

$W^j_{k,1}$ (four for each structure—spectral correlation j or attribute j that is taken into account) that makes it possible to calculate the similarity coefficient between the unknown structure X (1,0,0 ... x_j ... ... 0,1,0, x_k) and reference R (0,0,1 ... r_j ... 1,1,0, r_k) in a very efficient way:

$$S_{\mathbf{X},\mathbf{R}} = \sum_{j=1}^{k} W^j_{x_j r_j} \tag{4}$$

The similarity coefficient $S_{\mathbf{X},\mathbf{R}}$ may be called a weighted scalar product between two vectors X and R representing the unknown and reference spectrum—structure correlations, respectively. The components x_j and r_j of both vectors X and R are ones and zeros, depending on the presence or absence of a particular attribute j. The weights $W^j_{k,l}$ must be determined empirically by trial and error for each particular attribute, and this is not an easy task. The sum $S_{\mathbf{X},\mathbf{R}}$ is calculated for each compound in the library file. Eventually, as in the previously mentioned case, the structure of the unknown compound is claimed to be most similar to (or identical with) the structure of the compound that yields the best rating (largest sum).

CORRELATION OF SPECTRAL FEATURES WITH STRUCTURAL FRAGMENTS

The concept for the determination of structure developed by Gray [6] and Gribov et al. [3, 15, 16] may be called the most natural one, for it is very close to the scheme shown in Fig. 1. This approach to structure determination was first used in the construction of the DENDRAL system [17—19], and is done in two cycles: first, a goodlist and a badlist of all possible and forbidden structural fragments are formed; then, on the basis of these lists, molecular

formulae and chemical connectivity rules for the fragments, a structure is constructed which fits all the estimated requirements. The basic information for such a system comprises the sets of fragments containing complete structural and connectivity properties, together with all corresponding features for each type of spectrometry. Considering that the number of possible structures increases proportionally to $n!$ with increasing number of possible fragments n, it is obvious that such a method can be successful only for relatively small molecules containing up to about 20 non-hydrogen atoms.

Whilst the quality of the first part (the goodlist and badlist) depends very largely on knowledge of spectrum—structure relations, the efficiency of the second part (building the structure from the fragments) depends rather on the flexibility and quality of programming. Besides the discrepancy in the numbers of possible fragments used (which may easily be increased), the main difference between Gray's system and Gribov's system lies in the manner of testing the choice of fragments and their connectivity. In Gray's system, the "assign-fragment" routine, which has the form [6]

if: IR (data set$_1$, data set$_2$, ...) and/or
H-NMR (data set$_1$, ...) and/or
UV (data set$_1$, ...) and/or
MS (data set$_1$, ...)
then: - mark fragment accountable,
- create valence description,
- assign atoms,
- assign double bonds,
- correct molecular formula,

runs through all the possible conditions incorporated for all fragments; then the input spectra are no longer taken into account and a fragment can be removed from further consideration only if it does not fit the connectivity or combinatorial requirements.

In contrast, Gribov's system, which is based primarily on the analysis of infrared spectra [27], extracts possible fragments and structures in the first part only on the basis of vibrational analysis. For each hypothetical structure, a theoretical infrared spectrum is calculated and compared with the input spectrum. If the discrepancy is too large, the structure is rejected. After the list of hypothetical structures with calculated spectra close enough to the experimental one is complete, other spectroscopic methods, i.e. u.v., H-n.m.r., and mass spectrometry, become engaged in a refinement system. For each refined structure, a new vibrational spectrum is calculated and compared with the initial spectrum and the final ranking is made on the basis of this comparison.

On-line systems offer a completely different approach to the problem, enabling the user to influence the flow of the structural analysis directly during the process. The user can build up the proposed fragments or substructures and then change, combine or delete some of them, according to the response of the checks made for these hypothetical structures on the chosen file of spectra. To illustrate this very sensitive and practical method, an example based on the

NIH/EPA substructure program [9—11] and ^{13}C-n.m.r. file will be worked through [28, 29].

Figure 5 shows the stepwise construction of the methyl vinyl ketone skeleton based on simple commands as "CHAIN 4", "SATOM 5", "SBOND 2 5", etc. A careful user will notice that the structure built could be claimed as methyl vinyl ketone but could also be found as a fragment substituted on either or both sides of the structures of more complex compounds. At the end of the substructure search through a file of 4000 connection tables of the ^{13}C-n.m.r. file [30, 31], 50 compounds containing the specified query structure were found. The procedure was continued by checking the ^{13}C-n.m.r. spectra of all the 50 compounds retrieved. The results of this check, as displayed at the console, are shown in Fig. 6.

The distribution of the shifts of the methyl-C atom shown in the histogram on Fig. 6 is very broad and exhibits three distinct peaks; this suggests strongly

```
NIH/EPA

SUBSTRUCTURE SEARCH SYSTEM

FILE NOW CNMR

OPTION? CHAIN 4
OPTION? ABRAN 1 AT 2
OPTION? D

1??2??3??4
   ?
   ?
   5

OPTION? SATOM 5
SPECIFY ELEMENT SYMBOL = O

OPTION? SBOND 2 5 3 4
BOND TYPE (H FOR HELP) = CD

OPTION? SBOND 1 2 2 3
BOND TYPE (H FOR HELP) = CS

OPTION? D

1**2**3++4
   +
   +
   5O

OPTION?
```

Fig. 5. Construction of the methyl vinyl ketone skeleton by substructure search commands. If the atom type is not specified, it is assumed to be a carbon. The bond type must be specified (CS, single chain; CD, double chain, etc.) otherwise all types of bond (marked with ?) are considered as possible.

```
.CNMR
OPTION: SUB
TYPE THE NUMBER OF ATOM YOU WANT C NMR SHIFTS FOR: 1

THERE WERE 50 COMPOUNDS WITH A GIVEN ATOM.
WITH ASSIGNMENTS: 36, NOT OR PARTIAL ASSIGNED: 14
AMONG UNASSIGNED 12 WITHOUT THE REQUESTED SHIFT
WHICH ONES DO YOU WANT TO SEE?
TYPE OUT TO EXIT, OTHERWISE A (ASSIGNED) OR U (UNASSIGNED): A

  3:CAS REG.          141797 SHIFT  27.5 PPM
  4:CAS REG.          504201 SHIFT 125.6 PPM
  7:CAS REG.          684946 SHIFT  29.7 PPM
    .
    .
    .
49:CAS REG.        57031604 SHIFT  27.8 PPM

STATISTICS FOR 36 SHIFTS:
MAIN VALUE    : 36.5 PPM
ST. DEVIAT.   :  2.8 PPM
MIN. VALUE    : 24.3 PPM
MAX. VALUE    :125.6 PPM
TYPE N IF YOU DON'T WANT THE HISTOGRAM (Y1): Y2

 25 (PPM),FREQ.: 1 *
 27 (PPM),FREQ.: 7 *******
 29 (PPM),FREQ.: 6 ******
 31 (PPM),FREQ.: 2 **
 33 (PPM),FREQ.: 7 *******
 35 (PPM),FREQ.: 2 **
 39 (PPM),FREQ.: 3 ***
 43 (PPM),FREQ.: 5 *****
 51 (PPM),FREQ.: 2 **
125 (PPM),FREQ.: 1 *
```

Fig. 6. Statistical representation of chemical shift distribution calculated and displayed by the ^{13}C-n.m.r. option [28, 29] of the NIH/EPA system [2, 8]. The number of the C atoms for which C-n.m.r. shifts are required must be the same as assigned by the previous procedure (Fig. 5) from substructure search commands.

that at least three disparate types of carbon atoms have been included in this particular retrieval. The reason for this is that the vicinity of carbon atom number 1 was never specified apart from its being adjacent to a carbonyl group. To correct this shortcoming, the command "TERMA 1 1" was used to limit the links of carbon atom 1 to only 1 neighbor. When the entire procedure shown on Figs. 5 and 6 was repeated, the more reliable histogram shown in Fig. 7 was obtained. Because of the greater restrictions, the number of compounds retrieved was reduced from 50 to 22. Each carbon atom in the query structure could be treated in the same way as shown in Fig. 7, and four generalized shifts could be determined. The hypothetical structure is confirmed if

```
.CNMR
OPTION: SUB
TYPE THE NUMBER OF ATOM YOU WANT C NMR SHIFTS FOR: 1

THERE WERE 22 COMPOUNDS WITH A GIVEN ATOM.
WITH ASSIGNMENTS: 14, NOT OR PARTIAL ASSIGNED: 8
AMONG UNASSIGNED 7 WITHOUT THE REQUESTED SHIFT
WHICH ONES DO YOU WANT TO SEE?
TYPE OUT TO EXIT, OTHERWISE A (ASSIGNED) OR U (UNASSIGNED): A

 3:CAS REG.          141797 SHIFT 27.5 PPM
 7:CAS REG.         1522209 SHIFT 24.3 PPM
  .
  .
  .
22:CAS REG.        57031604 SHIFT 27.8 PPM

STATISTICS FOR 14 SHIFTS:
MAIN VALUE     :28.1 PPM
ST. DEVIAT.    : 0.5 PPM
MIN. VALUE     :24.3 PPM
MAX. VALUE     :32.3 PPM
TYPE N IF YOU DON'T WANT THE HISTOGRAM (Y1): Y2

25 (PPM),FREQ.: 1 *
27 (PPM),FREQ.: 7 *******
29 (PPM),FREQ.: 5 *****
33 (PPM),FREQ.: 1 *
```

Fig. 7. The same as Fig. 6 after the ambience of the methyl-C atom has been defined.

the main shifts obtained (with reasonable standard deviations) agree well with the experimental spectrum. The procedure explained may be repeated as many times as the user desires.

CODING OF CHEMICAL STRUCTURES

The problem is how to encode the chemical structure in an exact and canonical manner that will yield a notation easy to manipulate and acceptable to the computer [32—40]. The notation should fit many requirements; some of the most important criteria are that it should be unique for any structure, easy to encode and decode, economic in terms of computer memory, easy to manipulate in terms of substructures and fragments, and easy to grasp visually. Even the best notations do not fulfil all these requirements.

Connection tables [32] provide probably the best structural descriptions, are very convenient for substructure manipulation and are relatively easy to encode and decode, but they are very space-consuming and are slow in retrieval. In contrast, the Wiswesser Line-Formula Chemical Notation (WLN) [33] is easily understood, concise, and very good for problems such as retrieval of the

information whether the complete structure is precisely defined in the file or not; however, it is difficult for encoding or fragment manipulation. A structural fragment that can be clearly encoded in one structure may be completely hidden in another structure [40] (Fig. 8).

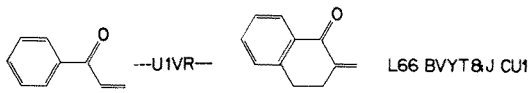


Fig. 8. Topologically equivalent fragments can be encoded quite differently in WLN.

In addition to complete notations, there are many fragment-coding systems (see, e.g., [41, 42]) that can be quite useful in particular cases [43, 44], but they do not satisfy the most important criterion, i.e. the unique description of each structure.

OTHER ASPECTS

Decision-making processes

In the last 10 years, the study of decision-making algorithms on the basis of pattern recognition has become very popular [45]. Although this field of research seems to be very exciting and promising at first sight, yet the excessive optimism of the early years has not, for the most part, been justified by the results [46]. The focus in research on learning algorithms (especially for chemical purposes) is now sliding from non-parametric towards parametric training methods [47]. There are some recent reports [48—50] on statistical (parametric) evaluation of "average spectra" as representations of specific structure—spectra correlations, which yield fairly good decisions. Figure 9 shows a hierarchal decision tree based on the average spectrum for each point of the tree. The upper part of Fig. 9 shows the average spectrum made for the first decision point. The problem with decision-making algorithms lies in the fact that it is relatively easy to achieve good predictive ability with comparatively small training and test sets, but if the number of points in the measurement space reaches the thousands or more, the learning procedure is either too time-consuming or gives predictive abilities too poor for serious consideration in information systems. A solution to the problem of how to extract better features from the patterns or measurement vectors would trigger very significant progress in this field.

The problem of mixtures

The problem of mixtures or multicomponent analysis is perhaps the most critical problem in automated structural elucidation and, indeed, in automated analysis of any kind. When the input spectrum represents completely unknown material, it is usually assumed that it pertains to a single compound, although in practice this is rarely the case. One of the first attacks on this problem was made by Sebasta and Johnson [51], who set reasonable basic requirements

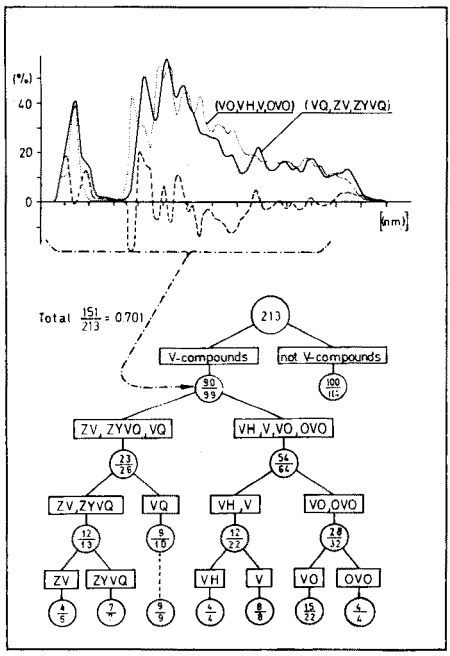


Fig. 9. Decision tree based on average spectra. The classes are coded by WLN (V: >C=O; VH: HC=O; VO: —C(=O)—O—, VQ: OH—C(=O)—, ZV: NH_2—C(=O)—; ZYVQ: NH_2—CH(—)—C(=O)—OH). Numbers in circles represent the ratio of correctly classified compounds.

for multicomponent analysis which were implemented in the MIRET (Multicomponent Infrared Retrieval) system. Isenhour [52] described a very fast algorithm designed for x-ray analysis of mixtures. The algorithm can be easily adapted for many other spectrometric methods. With a strategy slightly modified from that of Sebasta and Johnson, the MIXTURE option of the retrieval system ZAPAH2 [25, 53] has given some encouraging results in the elucidation of infrared spectra of real mixtures.

Some other trials [54] have been made in this direction, but the problem of mixtures is still without solution, and no combined system can deal with this problem at the present time [3].

Input and output of data

A further problem is the method used for input and output. There are normally very few problems with on-line systems: either the spectrometers are linked to the computer or the user types the data in and answers the displayed questions, or both. In either case, the flow of data is preprogrammed and the user does not need to learn much about the computer. The simple commands that are necessary for work must be displayed in such a way that any possible

response from the user is displayed. Slightly more work is required for batch-operating systems, where free-format input of all data must be provided for the user, because many chemists retain some distrust of computers.

Much more important than the input is the way that the final or intermediate results are output (displayed). A bad, bulky, incomplete and insufficiently informative output can disturb the user, and many valuable results or hints may be lost. The output design depends to a great extent on the hardware facilities available to the programmer and user. The output medium may vary from printed results on simple teletype or line printer to structural images on a colour display monitor [55]. Often it is desirable that the user should have some influence on the method and organization of the output.

CONCLUSION

The comparison of some combined chemical information systems makes it possible to judge their capabilities and deficiences. The efficiency of these systems is definitely inferior to the ability of expert analytical spectroscopists and even inferior to the interpretative faculty of an average student. What these systems can do really well is a fast search through large data files or collections which retrieves answers within prescribed or built-in tolerance limits. Evaluation of the answers retrieved is already very subjective and the source of many errors; real structure and substructure manipulation and decision-making processes based on spectral features can be made with tolerable performance only for a few very specialized problems.

The present state of the art in computerized complex chemical information systems cannot be properly considered as a matter simply of putting the final touches to an almost completed edifice ready for universal use, but rather as relatively small-scale pilot projects. Many ideas and methods have been tested; much more has yet to be done before the best strategies and algorithms can be chosen. But the field is large and the demand is great, and advances towards better and more complex on-line systems combining literature and experimental data, automatic decisions and other features will certainly come.

The author is greatly indebted to Prof. D. Hadži and Dr. J. T. Clerc for many helpful and fruitful discussions. The work was financially supported by the Research Community of Slovenia.

REFERENCES

1 D. Hadži, Computers in Chemical Research and Education, Vol. 1, Elsevier, Amsterdam, 1973.
2 S. R. Heller, G. W. A. Milne and R. J. Feldmann, Science, 195 (1977) 253, and The NIH-EPA Chemical Information System, Status Report No. 6, December 1977.
3 L. A. Gribov, M. E. Elyashberg and V. V. Serov, Anal. Chim. Acta, 95 (1977) 75.
4 J. T. Clerc and J. Zupan, Pure Appl. Chem., 49 (1977) 1827.
5 J. Zupan, M. Penca, D. Hadži and J. Marsel, Anal. Chem., 49 (1977) 2141.

6 N. A. B. Gray, Anal. Chem., 47 (1975) 2426.
7 V. A. Koptjug, Z. Chem., 15 (1975) 41.
8 P. R. Naegeli and J. T. Clerc, Anal. Chem., 45 (1974) 739A.
9 R. J. Feldmann in W. T. Wipke, S. R. Heller, R. J. Feldmann and E. Hyde (Eds.), Computer Representation and Manipulation of Chemical Information, J. Wiley, New York, 1974, p. 55.
10 R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller and B. Koch, J. Chem. Inf. Comp. Sci., 17 (1977) 157.
11 J. A. Miller, Substructure Search System, Users Manual (1977), Fein-Marquart, 7215 York Road, Baltimore, Md., 21212.
12 F. Erni and J. T. Clerc, Helv. Chim. Acta, 55 (1972) 489.
13 J. T. Clerc and F. Erni, Topics in Current Chemistry, 39 (1973) 91.
14 J. T. Clerc in D. Hadži (Ed.), Computers in Chemical Research and Education, Elsevier, Amsterdam, Vol. 2, 1973, p. 3/109.
15 L. A. Gribov, V. A. Dementyev, M. E. Elyashberg and E. Z. Yakupov, J. Mol. Struct., 22 (1974) 161.
16 L. A. Gribov in D. Hadži (Ed.), Computers in Chemical Research and Education, Elsevier, Amsterdam, Vol. 2, 1973, p. 3/81.
17 J. Ledreberg, G. L. Sutherland, B. E. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield and C. Djerassi, J. Am. Chem. Soc., 91 (1969) 2973.
18 D. H. Smith, B. G. Buchanan, R. S. Engelmore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg and C. Djerassi, J. Am. Chem. Soc., 94 (1972) 5962.
19 D. H. Smith, L. M. Masinter and N. S. Sridharon in W. T. Wipke, S. R. Heller, R. J. Feldmann and E. Hyde (Eds.), Computer Representation and Manipulation of Chemical Information, J. Wiley, New York, 1974, p. 287.
20 K. E. Iverson, A Programming Language, J. Wiley, New York, 1962.
21 D. E. Knuth, The Art of Computer Programming, Vol. 3, Sorting and Searching, Addison-Wesley, Reading, 1975, pp. 389, 506.
22 F. E. Lytle, Anal. Chem., 42 (1970) 355.
23 V. Y. Lum, P. S. T. Yven and M. Dodd, CACM, 14 (1971) 228.
24 W. Feller, An Introduction to Probability Theory, J. Wiley, New York, 1950, Section 2.3.
25 J. Zupan, D. Hadži and M. Penca, Comput. Chem., 1 (1976) 71.
26 J. Zupan and D. Hadži in E. V. Ludena and F. Brito (Eds.), Computers in Chemical Research, Education and Technology, Advances Studies Center IVIC, Caracas, 1977.
27 L. A. Gribov, J. Mol. Struct., 22 (1974) 353.
28 G. W. A. Milne, J. Zupan, S. R. Heller and J. A. Miller, Org. Mag. Reson., (1978) in press.
29 J. Zupan, S. R. Heller, G. W. A. Milne and J. A. Miller, Anal. Chim. Acta, 103 (1978) 141.
30 CNMR Data Base, NIC, Delft, PO Box 36 2600 AA, Delft, The Netherlands.
31 D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, Org. Mag. Reson., (1978) in press.
32 For Connection Tables, see, e.g.: H. L. Morgan, J. Chem. Doc., 5 (1965) 107, and references therein.
33 E. G. Smith and P. A. Baker, The Wiswesser Line-Formula Chemical Notation, 3rd edn., CIMI, Cherry Hill, N.Y., 1975.
34 C. M. Bowman, F. A. Laudel, N. W. Lee and M. H. Reslock, J. Chem. Doc., 8 (1968) 133.
35 C. M. Bowman, F. A. Laudel, N. W. Lee, M. H. Reslock and B. P. Smith, J. Chem. Doc., 10 (1970) 50.
36 J. E. Dubois in W. T. Wipke, S. R. Heller, R. J. Feldmann and E. Hyde (Eds.), Computer Representation and Manipulation of Chemical Information, J. Wiley, New York, 1974, p. 239.
37 W. T. Wipke and T. M. Dyott, J. Am. Chem. Soc., 96 (1974) 4834.
38 K. K. Agrawal, Transformation and Canonization Algorithms for Graph Representable Structures with Applications to a Heuristic Program for the Synthesis of Organic Molecules, Thesis, 1976, State University of New York at Stony Brook, Technical Report 63.
39 R. H. Bovie, Technical Report 70, (1977), State University of New York at Stony Brook, N.A. 11794.

40 J. Meili, Beitrag zur Automatisation im organisch-analytischen Laboratorium. Computerunterstüzte Dokumentation Spektroskopischer Daten, Thesis, Diss. ETH 5521, Zürich 1975, p. 144.
41 Codes and Instructions for WYANDOTTE-ASTM, American Society for Testing Materials, 1916 Race St., Philadelphia, Pa., 1964.
42 Handbook of CIDS Chemical Search Keys, Fein-Marquart Assoc., Baltimore, Md., 21212, (1973).
43 D. S. Erley, Anal. Chem., 40 (1968) 894.
44 J. Zupan, S. R. Heller and G. W. A. Milne, Vestnik SKD, 25 (1978) 37.
45 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, J. Wiley, New York, 1975, and references therein.
46 H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Bowie, G. A. Spritzer and J. E. Searleman, Science, 197 (1977) 1041.
47 N. Y. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
48 J. M. Comerford, P. G. Anderson, W. H. Snyder and H. S. Kimmel, Spectrochim. Acta Part A, 33 (1977) 651.
49 M. Penca, J. Zupan and D. Hadži, Anal. Chim. Acta, 95 (1977) 3.
50 C. L. Wilkins and T. R. Brunner, Anal. Chem., 49 (1977) 2136.
51 R. W. Sebasta and G. G. Johnson, Jr., Anal. Chem., 44 (1972) 260.
52 T. L. Isenhour, Anal. Chem., 45 (1973) 2153.
53 J. Zupan, J. T. Clerc and D. Hadži, Vestnik SKD, 23 (1976) 73.
54 C. E. Rechsteiner, H. S. Gold and R. P. Buck, Anal. Chim. Acta, 95 (1977) 51, and references therein.
55 R. J. Feldmann and T. K. Porter, Surface Representation of Biological Macromolecules, DCRT, NIH, Bethesda, Md. 20014.

Analytica Chimica Acta, 103 (1978) 289—302
Computer Techniques and Optimization

SISCOM — A NEW LIBRARY SEARCH SYSTEM FOR MASS SPECTRA

H. DAMEN, D. HENNEBERG* and B. WEIMANN
Max-Planck-Institut für Kohlenforschung, D-4330 Mülheim/Ruhr (West Germany)
(Received 26th May 1978)

SUMMARY

SISCOM is a library search system for mass spectrometry which is based on a new method of coding spectra by selecting the most important peaks within homologous ion series, and on a multiple factor assessment of the result. Examples demonstrate the ability of the system to identify various compounds, even from mixtures or by reference spectra which differ from those measured. SISCOM is especially suitable for detecting structural similarities like common substructures, even in cases where no similarity can be recognized by visual comparison of patterns or by human interpretation of the spectrum.

Library search methods generally operate in the following way: a matching procedure compares the unknown spectrum successively with the spectra of the library, for which purpose the spectral information has to be condensed by coding; then for each match a similarity or match factor is calculated and a ranked list of the n "best" matches is presented. The various search systems that have been reported differ in the kind of coding and matching. Regarding the coding, one group uses formal methods, e.g. the selection of all peaks from the whole spectrum which exceed a relative intensity threshold [1] or the n largest peaks from sections [2]. Others [3, 4, 5] apply more specific coding, i.e. attempts are made to select or derive the particular information which is relevant to interpretation of the spectra. The STIRS method [4] uses several different, specific codings in parallel. The matching can proceed in one or more steps (screening) [2].

Normally, the result of a search is assessed as a single number, called the match or similarity index or factor. The more sophisticated STIRS procedure [4] uses several match factors, each resulting from one of the different types of coding. The various methods have been developed with regard not only to the quality of the results attainable but also to the required storage and/or computing time [6]. The system developed in this laboratory is based on a new type of specific coding and a multiple factor assessment of the match results.

A library search can be carried out for different purposes. In biological or medical applications, the aim may be the identification of a compound known to be a typical component for the analysis in question and known to be represented in the library by a spectrum, the pattern of which will be very close

to the pattern obtained in the equipment used for the analysis. This type of library search can usually be carried out very quickly with a small, specialized library. Since the unknown may be impure or in a mixture, reverse search methods [7] can be applied, the result of which is more or less independent of the presence of additional components in the spectrum of the unknown. Another type of library search — encountered normally in service and research laboratories — has to deal with a wide range of compounds, the spectra of which may or may not be represented in the library. In this case, the library must be as comprehensive as possible. Of course, this means that the reference spectra are of very different origin (type of instrument, parameters of measurement, g.c.-m.s., direct inlet, etc.), and so the pattern of a given reference spectrum may differ from that of the measured spectrum.

The search system described here was developed for the second kind of analytical work. The most important typical situations are outlined in order of increasing difficulty where the unknown is:
(1) included in the library with a pattern close to the measured pattern;
(2) included in the library, but with a different pattern;
(3) a mixture spectrum, the components of which are included in the library;
(4) not included in the library, but where an interpretation is still possible, because similar compounds (homologues, isomers, compounds with the same partial structures or functional groups) are included and this similarity is reflected in some way in the spectrum and recognized by the search method.
These four situations may be viewed as different degrees of similarity. The greatest benefit from the use of a library search system is obtained if not only the first two, but all four, situations are handled satisfactorily.

The search system described here can handle the four situations; therefore the examples presented have been chosen primarily with emphasis on the more difficult cases rather than to demonstrate the ability of the system to identify different kinds of compounds. Because this search system is characterized by its ability to recognize different kinds and degrees of SImilarity, by using a Specific type of COding and Multiple matching factor assessment, it has been called SISCOM.

THE SEARCH SYSTEM

Coding of spectra

Coded spectra consist of selected mass numbers (characteristics) and their corresponding intensities. The mass number of a peak in the spectrum is regarded as a characteristic if two conditions are met: first, the isotope-corrected intensity must exceed the arithmetic mean of the higher and the lower homologous neighbours; secondly, the intensity of the peak must be significant, i.e., greater than a relative threshold (normally 2%) and, in the measured spectrum, also greater than some absolute value which depends on the signal-to-noise ratio. The m/e 28 and 32 peaks are not allowed as characteristics. Special conditions are defined for m/e 42 (undefined lower

homologous neighbour) and *m/e* 44 (a background peak that cannot be correlated). The full spectra in Figs. 1 and 8 with the marked characteristics can be taken as an example of the coding method used.

Coding procedures described in the literature select either all peaks with relative intensities above a constant threshold or the *n* largest peaks in consecutive intervals of length *m*, e.g. 2 out of 14. This means that peaks may be included as characteristics which are not specific for a structure, or specific peaks may be excluded. For instance, many of the intense ions in the lower mass range of paraffin-like spectra do not provide selective information. However, highly characteristic ions are sometimes observed in the neighbourhood of more intense fragments and may therefore be suppressed by the coding procedure. SISCOM extracts the ions exceeding defined limits not from a neighbourhood on the mass scale, but from the neighbouring homologous ions. This type of coding selects peaks, including small ones, which are relevant for mass spectrometric interpretation.

Matching procedure

The matching procedure is based on the following six comparison factors:
N_C : the number (N) of common (C) characteristics found in the sample spectrum and reference;
N_R : the number (N) of characteristics remaining in the reference (R);
N_S : the number (N) of characteristics remaining in the sample spectrum (S),
I_R : the sum of the intensities (I) of the characteristics in N_R, relative to the sum of the intensities of all characteristics in the reference (R);
I_S : as I_R, but for N_S and the sample spectrum (S), respectively;
P_C : correspondence of the pattern (P) of the common (C) characteristics in the sample spectrum and reference, given as a modified correlation coefficient (identity 100, upper range extended).

To illustrate the coding and matching procedure, a spectrum (unknown) is compared with a single reference spectrum in Fig. 1. The three types of characteristics are marked; the resulting values of the comparison factors are noted in the legend. This example is a simple case of situation 4: a partial structure common to the unknown and the reference compound gives rise to an easily recognizable and very similar partial spectrum. Thus, a good idea of the unknown structure can be obtained even when the unknown compound is not in the library. In general, however, the relation between (partial) spectrum and (partial) structure is not reversible, can be ambiguous and is sometimes unrecognizable even when present.

The ranking of matches

The ranking process is accomplished in two distinct steps. The three comparison factors N_C, N_R and N_S are used in the first step to calculate a value B, on the basis of which the n (normally 150) best matches are determined.

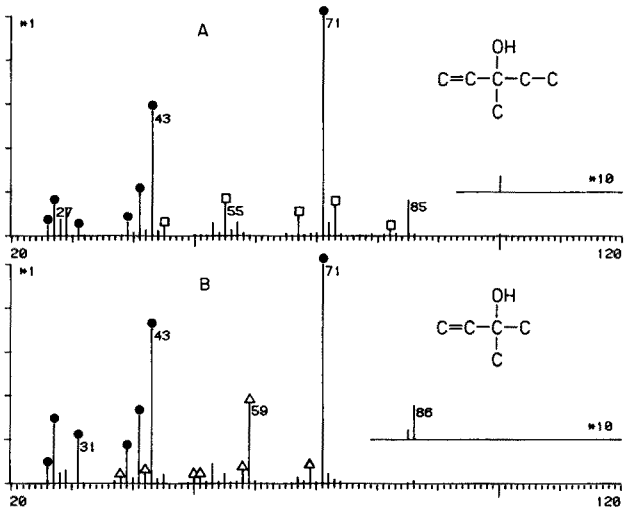


Fig. 1. Coded spectrum of the "unknown" (A) and of the reference (B). The three types of characteristics are marked as follows: common (•), remaining in the unknown spectrum (□), and remaining in the reference spectrum (△). The resulting comparison factors are: $N_C = 7$, $N_R = 7$, $N_S = 5$, $I_R = 17\%$, $I_S = 18\%$, $P_C = 87$.

$$B = N_C/(a*N_R + b*N_S + c) \quad (1)$$

where a, b and c are empirical weighting factors. For these 150 best reference spectra, a combination of all comparison factors is then used in a second step, to establish an optimal sequence from the equation

$$S = F1(N_C) + F2(P_C) + F3(N_R) + F4(I_R, I_S) \quad (2)$$

where $F1$—$F4$ are empirical weighting functions. S can thus be viewed as a measure of "similarity" between the unknown and the corresponding reference spectrum. The use of comparison factors like I_R, I_S or P_C in the first step would increase the computing time appreciably, because this step is applied to all spectra in the library.

To evaluate the empirical factors and functions in eqns. (1) and (2) a test set of spectra containing two groups of spectra was used. One group consisted of library spectra of simple as well as complicated compounds with different functional or structural features. The second group contained spectra from routine analysis including mixtures and impure compounds. Further spectra were added to the second group when they were found to supply characteristic or new aspects to the interplay of the empirical factors and functions.

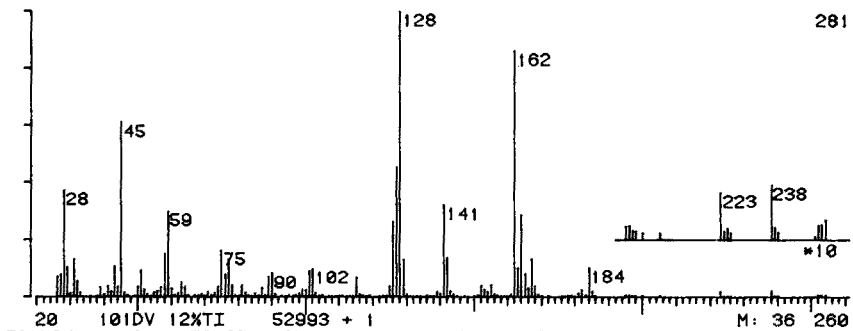
Output of results

The results of the search are output on display terminals (Tectronix 4012) together with a hardcopy unit. The information available for output consists of the unknown spectrum with label, number, comparison factors, formula and name of the 150 best reference spectra. If our own library was searched, the structures [8] of the selected references are also included.

The output of this information is arranged according to the capacity of the display as follows. The upper half of the screen is used to present the submitted unknown spectrum. The lower half contains a table with the information on the 16 best references. Figure 2 is an example of such a page of output. It is very important that the unknown spectrum be viewed together with the search result, because often a direct feedback between the spectra selected in the search and the unknown spectrum is needed for optimal interpretation.

The order of the 150 reference spectra obtained by means of eqn. (2) normally guarantees that the most important reference spectra for the search appear among the earliest, so that inspection of the later matches is unnecessary. If examination of the 16 best matches leaves any doubts or suggests that more information would be useful, the remaining matches can be presented page by page each with 35 reference spectra. When the internal library is used, the structures of the best matching references can be plotted instead of the table with names (see Fig. 9).

An important feature of SISCOM is that the different comparison factors are not only involved in selecting and ranking the best reference spectra, but are also presented in the output. The reason for this is that in any library application a single number — in SISCOM the value of S in eqn. (2) — is



EG1727	0	10	0:56	87	C10H7CL	1-CHLOR-NAPHTHALIN
EG2464	2	10	6:58	63	C10 H8	NAPHTHALIN
EG1100	7	14	11:82	0	C14H14	1,1-DIPHENYLETHANE
EG1330	5	9	14:78	65	C4H10O3	1-METHOXI-2,3-PROPANDIOLE
EG0243	40	24	59:35	3	C14H12O4	2-(1,4-DIOXAN-2-YL)-1,4-NAPHTHOCHINON
EG2070	3	9	2:65	0	C14H14	2,2'-BITOLOYL
EG1513	1	10	1:91	0	C6H6	BENZOL
EG2524	7	14	18:82	4	C8H7 BR	P-BROM-STYROL
EG2049	8	12	14:67	0	C14H14	P-METHYL-DIPHENYLMETHAN
EG1278	0	9	0:92	0	C6H6	BENZOL
EG2474	3	12	17:80	0	C12H10 O2 S	DIPHENYLSULFON
EG2552	8	11	13:74	0	C13H12 S	DIPHENYLMETHANTHIOL
EG3533	9	20	57:53	0	C12H8N2	PHENYL, 1-CYANO-, 2-(2-PYRIDYL)-
EG3228	13	12	74:57	85	C11H12	1-PHENYLCYCLOPENTEN
EG2071	3	9	10:69	0	C14H14	3,3'-BITOLOYL
EG3186	25	21	56:51	30	C16H18 O	1-(ALPHA-NAPHTHYL)-CYCLOHEXANOL-(1)

Fig. 2. Complete first page of the output of a search result with identification of mixture components. M:36 given just below the spectrum means that there are 36 characteristics in the spectrum. The values of N_S for each line can then be derived as $M-N_C$.

insufficient to characterize a match result. The functions in S are chosen as a compromise to give relatively good ranking for all of the four different situations mentioned. The output of all the comparison factors, however, enables a trained interpreter to identify the situation typical of the particular search, and then to make a refined assessment of the validity or meaning of the "best" reference spectra presented. The experience needed to make full use of the information output can be gained in a very short time by anyone who is familiar with the interpretation of mass spectra, if suitable sets of training examples are used.

RESULTS AND DISCUSSION

Data bases

The following four data bases are currently implemented with the system and can be used singly or combined arbitrarily:
(1) the Institute's collection of spectra [9], currently 3300, labelled E or EG;
(2) the EPA/NIH/MSDC 1977 collection with 25560 spectra, labelled B;
(3) the old EPA/NIH collection with 33600 spectra and 9000 doublets, labelled X; and
(4) a collection of 524 steroid spectra [10], labelled S.
The label for each library in combination with the spectrum number identifies or distinguishes the spectra in single or mixed data base searches.

The mean number of characteristics per spectrum is 19 for data base E and 24 or 25 for B, X and S. The storage required for 10000 condensed spectra together with names (one for each reference) and formulae is 1.5 MBytes, 45% of which are needed for names and formulae.

Search examples

The first example is a search in data base E; the result is given in Fig. 3. The spectrum of 2-octanol, submitted as unknown, was of poor quality, having been obtained from a g.c.—m.s. run with high column bleeding by background subtraction. In the uncorrected spectrum, 60% of the total ionization was

```
EG2744  2  9  4: 5 99 C8H18 O      2-OKTANOL
EG1061  6  8 20: 8 92 C12H26O      2-DODEKANOL
EG3216  4  7 18:20 98 C6H14O       2-HEXANOL
EG0941  4  6 19:22 99 C10H22O      2-DEKANOL
EG1183  3  5 11:17 97 C5H12O       2-PENTANOL
EG1891  3  4 14:29 ** C3H8O        ISOPROPANOL
EG2961  7  6 19:22 83 C6H13 O CL   2-METHYL-(RR)-3-CHLOR-(SS)-4-PENTANOL
EG3450  3  4 19:27 98 C5H10 O      4-PENTEN-2-OL
EG2132  5  7 13:11  1 C10H22O      DI-(2-PENTYL)-AETHER
EG1179  4  4 20:29 98 C4H10O       (-) BUTANOL-2
EG2376  4  5  8:70 78 C10H20       2-DECEN TRANS
EG2363  2  5 46:17 77 C7H16O       3-METHYL-3-HEXANOL
EG2044  6  6 13:10 24 C6H14O2      1,4-DIMETHOXIBUTAN
EG3499  3  7 10:63 26 C9H18        3-HEPTEN, 3-AETHYL-
EG1330  8  6 36:12 61 C4H10O3      1-METHOXI-2,3-PROPANDIOLE
EG2892  6  5  8:73 76 C10H18 O     1-DECIN-4-OL
```

Fig. 3. Result of a search for the identification of 2-octanol. The sixteen best reference spectra are listed with name, formula and comparison factors. Columns 1—6 represent: Label with spectrum number, N_R, N_C, $I_R : I_S$, P_C (** for 100).

squalane. Except for the P_C value of 100 (**) for isopropanol, all comparison factors shown in the first line for 2-octanol are the best in the particular columns, i.e., highest N_C and P_C, lowest N_R, I_R and I_S. This is a clear identification. The following lines show eight additional alcohols with the hydroxy group in the 2-position.

The next example (Fig. 4) is a search in data base X with the spectrum of farnesol (*cis—trans*), taken from the Registry of Mass Spectral Data [11] (No. AA-1578-1) as unknown. Again, the comparison factors for spectrum X10448 in the first line are the best and are very good, indicating an identification. In fact, the spectrum taken as unknown seems to be the same as spectrum 10448 in Data Base X; the very small deviations from absolute identity result from the fact that the spectrum was acquired from the graphical representation in the book and not from a digital tape. Among the 16 best matches in Fig. 4, there are seven farnesol spectra and five farnesol derivatives. In the remaining 134 best matches, there were five additional farnesol spectra and 95% of the rest were mono-, sesqui- and triterpenes. This shows that terpenes are recognized selectively, but that farnesol is distinctly preferred, although the reference spectra are of different origins. When the farnesol spectrum X22719, which has relatively bad comparison factors (P_C = 39) and which indeed differs remarkably from the unknown, was taken as unknown, farnesol was found three times, once in first position, among the 16 best matches. This example demonstrates the ability of SISCOM to deal with situations where the unknown compound is included in the library, but differs from the measured spectrum because of different instrumental parameters.

A further example, shown in Fig. 2, demonstrates how components of a mixture can be identified by interpretation of characteristic constellations of comparison factors. The highest N_C found among the 16 best matches is 24, but all other comparison factors for this reference have bad values. There are many more missing characteristics (N_R = 40) than common, the missing characteristics have appreciable intensity (I_R = 59%) and there is no correlation between the common characteristics (P_C = 3). Therefore this reference is not taken into account for interpretation of the spectrum. The comparison factors in the first line of the table, however, indicate that all 10 character-

X10448	1	21	0:	2	99	C15H260	FARNESOL CIS-TRANS
X22724	7	18	9:	7	91	C15H260	FARNESOL
X22726	9	17	10:10		91	C15H260	FARNESOL
X10449	3	13	2:14		97	C15H260	FARNESOL TRANS-TRANS
X22718	4	15	4:10		83	C15H260	FARNESOL
X12808	5	17	5:10		70	C17H2802	FARNESYL ACETATE
X21644	11	14	22:12		88	C10H180	LAVANDULOL
X22719	12	18	20:	7	39	C15H260	FARNESOL
X10287	10	14	9:10		72	C15H240	CIS,TRANS-FARNESAL
X10288	5	13	5:13		67	C15H240	TRANS,TRANS-FARNESAL
X 9210	6	15	11:11		60	C15H24	(Z)-BETA-FARNESENE
X22555	7	17	13:10		19	C15H24	ALPHA-FARNESENE
X22727	6	15	7:12		44	C15H260	NEROLIDOL ISOMER
X 9219	5	16	15:17		55	C15H24	BETA-BISABOLENE
X22725	12	18	18:42		64	C15H260	TRANS-TRANS-FARNESOL
X10447	8	16	9:11		36	C15H260	NEROLIDOL ISOMER

Fig. 4. Identification of farnesol in a data base containing doublets. The columns have the same significance as in Fig. 3.

istics of 1-chloronaphthalene are present in the unknown (N_R = 0) and the correlation of the common characteristics is good (P_C = 87). But the spectrum contains more characteristics, N_S = 26 (M—N_C, see legend to Fig. 2) and I_S = 56%. This is a typical constellation for the identification of a component of a mixture. The same is true with somewhat less significance for the second line, indicating naphthalene to be a second component present in the sample. The two entries of benzene with a similar constellation for N_R and N_C do not suggest benzene as a component, because P_C is zero in both cases (an adequate *m/e* 78 is missing from the spectrum). Also 1-phenylcyclopentene with N_C = 12 and P_C = 85 cannot be a component, because 13 of its characteristics with I_R = 74% are missing.

Examination of the spectrum makes it clear that chloronaphthalene and naphthalene are not the only components in the spectrum. Among the main ions still not assigned are *m/e* 45, 59, 75, 141 and 184. Further inspection of the comparison factors indicates that one of the best remaining matches is the fourth one: N_R and I_R are relatively small, while P_C is quite high. These values in combination with a moderate N_C value do not suggest very close similarity of a further component of the mixture with the corresponding reference 1-methoxy-2,3-propanediol. But the functional groups indicated are probable for the ions cited. Without going into detail, it should be mentioned that changing the search parameters or running a new search after subtraction of the data for chloronaphthalene and naphthalene, made it possible to identify monoglyme; butylnaphthalene was indicated as a further component of the mixture.

This discussion of a search result illustrates how the availability of all single comparison factors allows a refined and more substantiated assessment of the validity of the reference spectra offered as best matches. While this example can be characterized by the fact that the main peaks of the components of the mixture do not overlap severely, the next example represents a mixture with overlapping spectra.

Figure 5 shows a g.c.—m.s. run of a mixture of steroids with known composition, one peak of which is composed of strongly overlapping components, as shown by the mass chromatograms in the lower half. Spectrum 35 corrected for background (spectrum 27 subtracted) was taken as the unknown. The search was made in data base S, which is a collection of 524 steroids. The results (Fig. 6) show that the P_C values are generally low. By comparing reference spectra from data base S with the steroid spectra from the g.c.—m.s. run, it is evident that the spectra in S had been measured by the direct inlet method. Thus, the peaks in the reference spectra in the higher mass range, and the molecular peak in particular, are much more intense. This explains the low correlation of the common characteristics. Therefore the P_C values should not be used for evaluation of the matches. Further inspection of the comparison factors leads to the conclusion that the references up to S5075, with the exception of S5077, are good matches, because N_R is low and/or N_C is high. S5077 ends up in the high position by virtue of its high P_C which is

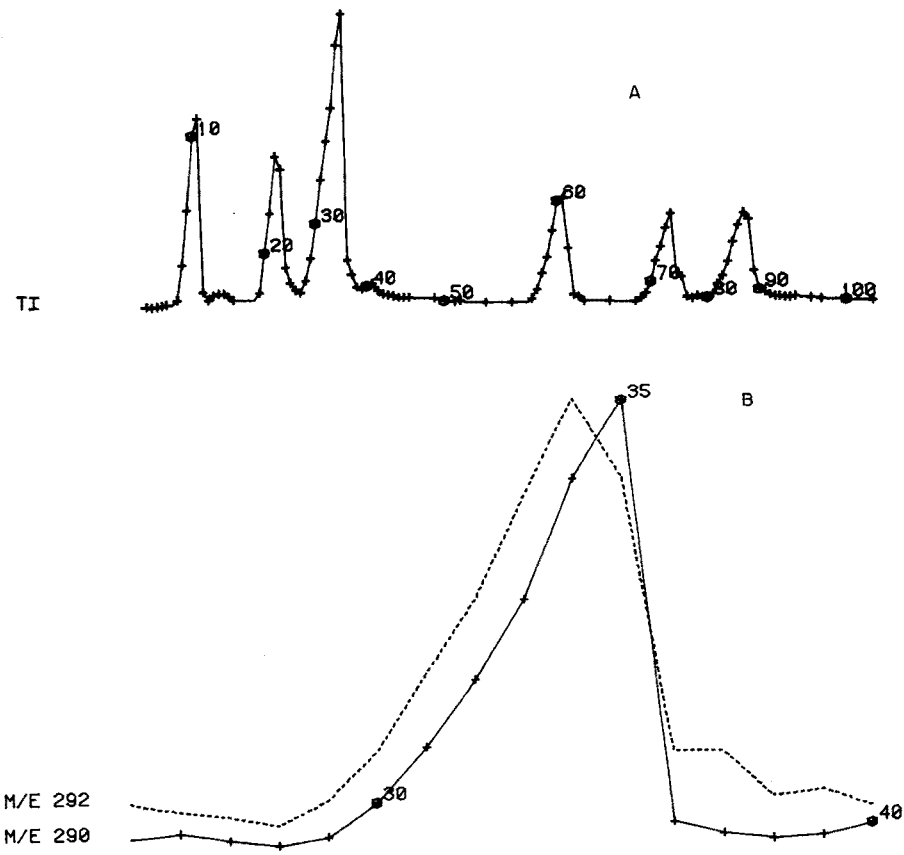


Fig. 5. G.c.—m.s. run of a mixture of underivatized steroids. (A) Reconstructed total ionization chromatogram with spectrum numbers; (B) section from spectrum (25—40), mass chromatograms, showing the overlap of two compounds with molecular masses *m/e* 290 and 292.

S	5103	8	37	12:18	12	C19H3002	17-BETA-HYDROXY-5-ALPHA-ANDROSTAN-3-ON
S	5077	18	33	40:22	78	C19H3203	5-ALPHA-ANDROSTAN-2-BETA,3-BETA,17-BETA-TRIOL
S	5105	5	30	6:25	7	C19H3002	17-ALPHA-HYDROXY-5-ALPHA-ANDROSTAN-3-ON
S	5064	8	31	12:28	7	C19H3202	5-ALPHA-ANDROSTAN-3-BETA,17-BETA-DIOL
S	5065	4	29	8:31	4	C19H3202	5-ALPHA-ANDROSTAN-3-BETA,17-ALPHA-DIOL
S	5063	9	32	13:31	11	C19H3202	3-ALPHA,17-BETA-DIHYDROXY-5-ALPHA-ANDROSTAN
S	5062	5	29	9:32	4	C19H3202	5-ALPHA-ANDROSTAN-3-ALPHA,17-ALPHA-DIOL
S	5075	14	37	23:31	25	C19H3202	3-BETA,17-BETA-DIHYDROXY-5-BETA-ANDROSTAN
S	5056	12	32	19:30	9	C20H340	17-BETA-HYDROXY-1-ALPHA-METHYL-5-ALPHA-ANDROSTAN
S	5104	23	36	26:27	17	C19H3002	17-BETA-HYDROXY-5-BETA-ANDROSTAN-3-ON
S	5478	11	30	18:30	7	C19H3203	5-ALPHA-ANDROSTAN-3-BETA,12-BETA,17-BETA-TRIOL
S	5093	20	33	23:24	3	C19H3002	3-BETA,17-BETA-DIHYDROXY-5-ALPHA-ANDROST-8(14)-EN
S	5399	19	31	19:31	20	C19H3002	3-BETA-HYDROXY-5-BETA-ANDROSTAN-17-ON
S	5414	23	33	29:29	46	C19H3203	5-BETA-ANDROSTAN-3-BETA,16-BETA,17-BETA-TRIOL
S	5161	36	33	47:25	55	C19H3003	3-ALPHA,6-ALPHA-DIHYDROXY-5-BETA-ANDROSTAN-17-ON
S	5071	14	29	17:27	6	C19H3203	5-ALPHA-ANDROSTAN-3-ALPHA,11-BETA,17-BETA-TRIOL

Fig. 6. Search results for a mixture of two steroids. One component (S5103) is identified, and the other (S5063) and stereoisomers are at the top of the table.

of no significance in this search. With regard to preferences within this group of seven reference spectra, the first one, S5103, is indeed identified as the best (best combination of high N_C with low N_R). Of the remaining six, only S5075 may be slightly preferable because of its high N_C value. In fact, reference S5103 is one of the two components in the mixture spectrum, S5075 is a stereoisomer, S5063 is the other component, and the other four reference spectra among the group of the seven best are stereoisomers of the second component. The search for this spectrum in the other data bases has, in principle, the same result, but reflects the content and the quality of the steroid spectra in the expected way.

The search results given in Figs. 3, 4 and 6 have included the selection of reference spectra with similar partial structures. The following example illustrates the selection of compounds with different degrees of structural similarity. The spectrum of a urea derivative — 3-(2-chloro-*p*-tolyl)-1,1-diethylurea — taken from data base B was used as unknown and searched in base B. The results (Fig. 7) show three groups of reference spectra among the 16 best matches. The first group (lines 2—4) is defined by low N_R and I_R and very high P_C values. The three compounds are very close isomers of the unknown with very similar spectra. The second group (lines 5—7 and 9) have relatively high N_C, medium N_R and no P_C values. This is again a well-defined group of ureas but with spectra which differ considerably from that of the unknown. The remaining reference spectra, except for benzene, are for aromatic nitrogen compounds, two of which have chlorine attached to the aromatic ring.

In the preceding examples, it seems possible to rationalize the similarities among the best-matching compounds by similarities in parts of the spectra, which can be discovered by a close inspection of the patterns. This is particularly well illustrated by the example in Fig. 1. However, during systematic tests of SISCOM in analytical practice over 18 months, several examples were encountered where those features of a spectrum responsible for the selection of references with great structural similarities to the unknown could not be detected by examination of the pattern. To conclude the search examples, two typical cases of the latter type will be presented.

```
B13771   0 28   0: 0 ** C12H17CLN2O UREA, 3-(2-CHLORO-P-TOLYL)-1,1-DIETHYL-
B13780   8 22   7:13 92 C12H17CLN2O UREA, 3-(5-CHLORO-O-TOLYL)-1,1-DIETHYL-
B13772   4 11   8:30 90 C12H17CLN2O UREA, 3-(4-CHLORO-O-TOLYL)-1,1-DIETHYL-
B13705   5 14  17:22 84 C12H17CLN2O UREA, 3-(3-CHLORO-O-TOLYL)-1,1-DIETHYL-
B18119  14 19  31:72  0 C15H14CL2N2OUREA, N,N'-BIS(2-CHLORO-6-METHYLPHENYL)-
B18117  10 16  27:74  0 C15H14CL2N2OUREA, N-(2-CHLORO-3-METHYLPHENYL)-N'-(3-CHLO
B18120  12 16  25:76  1 C15H14CL2N2OUREA, N,N'-BIS(5-CHLORO-2-METHYLPHENYL)-
B11517  13 12  17:82 46 C8H8N2SE    SELENOCYANIC ACID, 4-AMINO-O-TOLYL ESTER
B18118  10 14  28:76  0 C15H14CL2N2OUREA, N,N'-BIS(3-CHLORO-2-METHYLPHENYL)-
B 4072  13 12  21:87  6 C7H8CLN      P-TOLUIDINE, 2-CHLORO-
B  318   1  7   2:92  0 C6H6        BENZENE
B 7017   5 10  18:81  0 C9H10CLN     AZETIDINE, 1-CHLORO-2-PHENYL-
B 7957  10 11  51:85 72 C9H8N2O2     CARBOSTYRIL, 3-AMINO-1-HYDROXY-
B 3331  13 13  31:74  0 C9H11N       AZETIDINE, 2-PHENYL-
B 3454  10 10  46:90 72 C8H9NO       BENZALDEHYDE, O-METHYLOXIME
B11023  16 16  66:47  5 C12H18N2O    BENZAMIDE, N-BUTYL-2-(METHYLAMINO)-
```

Fig. 7. Search result for spectrum B13771 (first line) as unknown. Similar compounds are two groups of urea derivatives and aromatic nitrogen compounds.

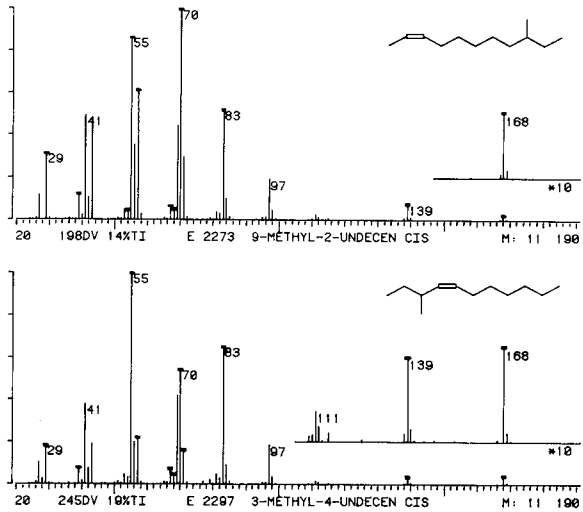


Fig. 8. Spectra of two isomeric methylundecenes with the characteristics marked: strong correspondence of characteristics, bad correlation of intensities, and similar partial structure.

Figure 8 shows the spectra of two compounds which by interpretation can be assigned as hydrocarbons of the formula $C_{12}H_{24}$. From the characteristics marked, it can be recognized that nearly all characteristics are common, although their pattern is very different. The question arises whether the high degree of correspondence in the characteristics reflects a similarity between the structures of the compounds. In the present state of knowledge of mass spectra of olefins, the 3-methyl branch is not considered to be this similarity, because this function is isolated in the 9-methyl-2-undecene and allylic to the double bond in 3-methyl-4-undecene. Figure 9 shows the result of a search in data base E as a facsimile of the output of structures instead of the usual table. It should be mentioned that data base E contains 190 mono-olefins (including all methyl-undecenes) and 140 cycloparaffins.

The result of the search is that the 3-methyl function is indeed a structural similarity selected by SISCOM in the mass spectra of olefins, independently of the position of the double bond and the molecular weight. This is a quite unexpected result. Another type of coding, based on the fourteen summed intensities of the homologous ion series (modulo 14 reduction [12]), gives an indication of the position of the double bond [13]. This shows that special types of coding make particular types of similarities recognizable. This has also been tested by using different match factors in the STIRS system [14].

The last example of a search originated from the analysis of a synthesized compound. Figure 10 shows the spectrum of the product ("unknown") together with the spectra of the three best matches. It is evident (also from the comparison factors which are not given here) that the best match identifies the

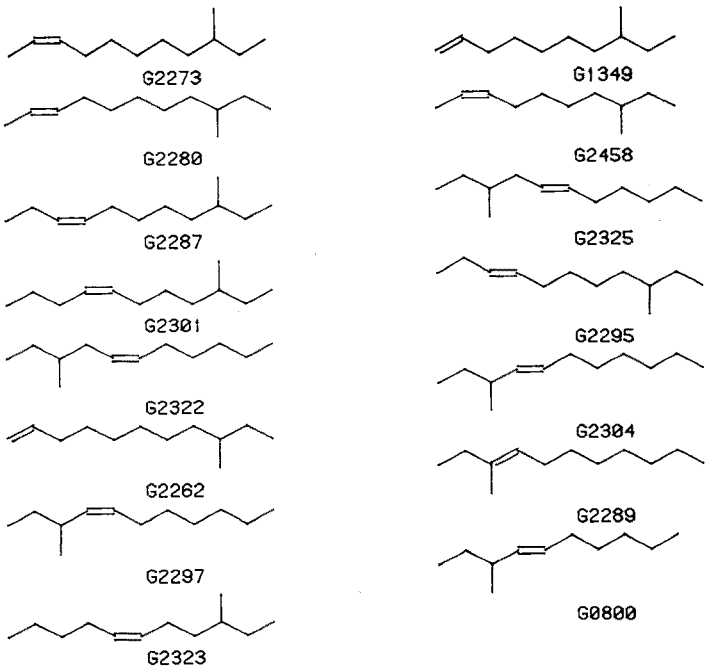


Fig. 9. Search result with the upper spectrum of Fig. 8 as unknown. Output of the best references is given in form of the structures (facsimile). The 3-methyl branch is shown as a common partial structure for C_{11} and C_{12} olefins, with varying position of the double bond.

product and also shows impurities. The given structures of the references show a clear similarity among the structures, although there is a striking dissimilarity between the patterns.

These last two examples were two of the cases found which suggested that a suitable search system can detect previously unknown relationships between structure and mass spectrum or relationships that cannot be detected by normal visual interpretation of patterns.

USE AND FURTHER DEVELOPMENT

SISCOM can be used in two ways. First, it can be activated in an interactive dialogue during evaluation of an analytical problem at the terminal. In this case, there is also the possibility of changing search parameters (see below). Secondly, the search can be coupled to a program [15] which automatically selects (and corrects for background in the case of g.c.—m.s.) the relevant spectra from a series measured by g.c.—m.s. or during a fractionating evaporation of solids. In this case, SISCOM is an integral part of a completely automated procedure of evaluating series of spectra.

The use of SISCOM has proved to be very helpful for typical problems in

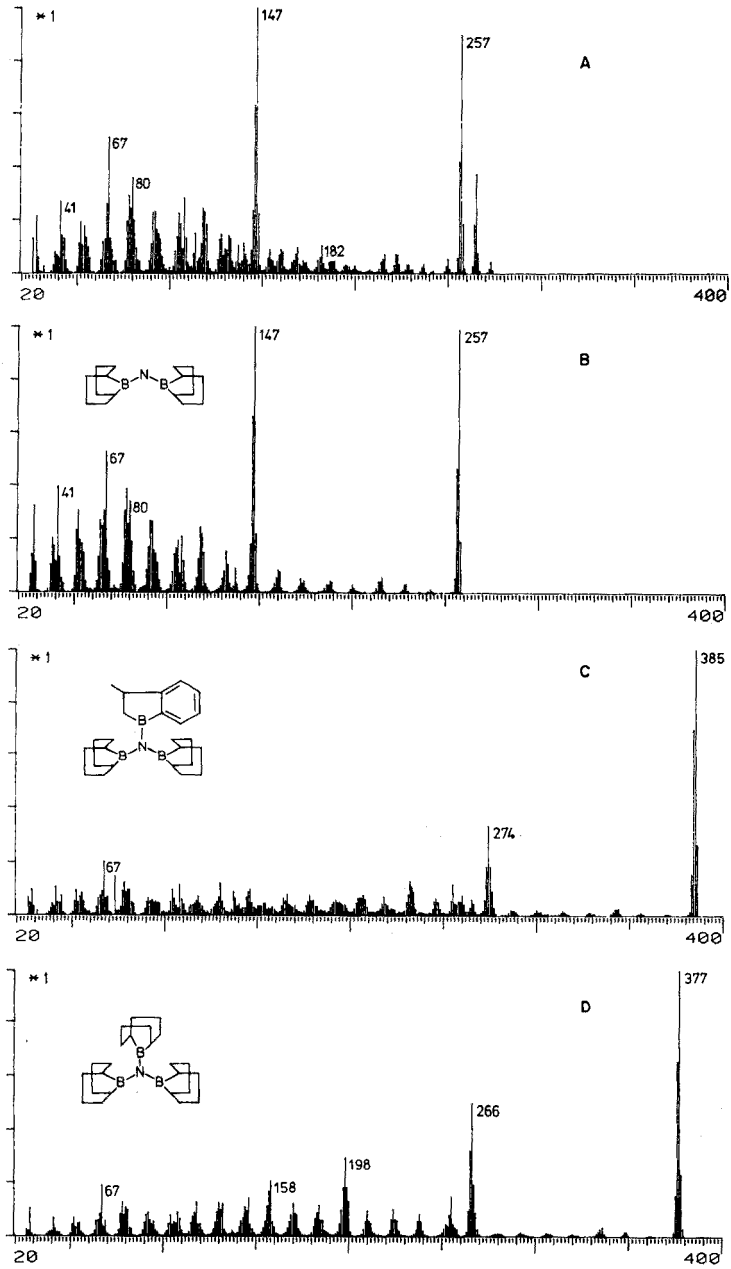


Fig. 10. Spectrum of the unknown (A) with the spectra of the three best matches (B—D). Search result: identification (first match) of the slightly contaminated compound, with identical partial structures in the reference compounds, but no similarity visible in the full pattern.

an analytical service laboratory. Many identifications have been made of compounds which were present in a library but were not known or not familiar to the interpreter, and many decisive hints have been given for compounds which were not included in any of the libraries. The time (CPU time) needed for a single complete search in the time-sharing computer center of the Institut (DEC-system 10) is 10—15 s per 10000 spectra.

Several systematic investigations have been carried out dealing with variations of several search parameters or conditions: (1) differences in results when data bases with different contents are searched; (2) the effect of mixing a more specialized with a comprehensive library; (3) the influence of using only characteristics in a limited mass range; (4) results obtained by using only characteristics that have intensities above a given threshold. These current investigations are part of a more general approach to improve the results of SISCOM, for instance, by changing the constants and functions in the selection and ranking of references according to the search situation concerned.

REFERENCES

1 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
2 H. S. Hertz, R. A. Hites, and K. Biemann, Anal. Chem., 43 (1971) 681.
3 P. R. Naegeli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
4 G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer, and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
5 K. S. Kwok, R. Venkataraghavan, and F. W. McLafferty, J. Am. Chem. Soc., 95 (1973) 4185.
6 L. E. Wangen, W. S. Woodward, and T. L. Isenhour, Anal. Chem., 43 (1971) 1605.
7 F. P. Abramson, Anal. Chem., 47 (1975) 45.
8 E. Ziegler and K. Boll, Anal. Chim. Acta, 103 (1978) 237.
9 Submitted for inclusion in the EPA/NIH Data Base.
10 Origin, G. Spiteller, University of Bayreuth, Germany; Compilation, K. Varmuza, Tech. University of Vienna, Austria.
11 Registry of Mass Spectral Data, Vols. 1—4, E. Stenhagen, S. Abrahamsson and F. W. McLafferty, J. Wiley, New York, 1974.
12 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1464.
13 D. Henneberg, unpublished work.
14 H. E. Dayringer, G. M. Pesyna, R. Venkataraghavan, and F. W. McLafferty, Org. Mass Spectrom., 11 (1976) 529.
15 D. Henneberg, H. Damen, and B. Weimann, in N. R. Daly (Ed.), Advances in Mass Spectrometry, Heyden, London, 1978, p. 975.

Analytica Chimica Acta, 103 (1978) 303—311
Computer Techniques and Optimization

DESIGN AND APPLICATION OF LOW-COST INFRARED DATA SYSTEMS

J. P. COATES
Perkin-Elmer Limited, Beaconsfield, Bucks. (Gt. Britain)
S. GEARY
Perkin-Elmer Corporation, Norwalk, CT (U.S.A.)

(Received 31st May 1978)

SUMMARY

The adaptation of a standard minicomputer to the processing of infrared spectroscopic data is described. The computer is interfaced to one of three high-performance infrared spectrophotometers and is supported by a flexible software package designed for the spectroscopist. The development of these computerized dispersive spectroscopy systems is discussed in terms of hardware and software. An indication of performance is included by reference to specific analytical applications.

Early attempts to acquire data in a computer-compatible form from a standard dispersive infrared instrument were based on obtaining an analog signal (0—1 V) from the recorder servo system. These were limited to off-line processing, usually from paper or magnetic tape medium. Digitization involved analog-to-digital (A/D) conversion of the output voltage. This was sampled at a rate defined by an external clock controlled by a crystal oscillator or by mains frequency. Certain instrument modifications were also required to synchronize the data logging to the spectrum scan, especially to accommodate instrument functions such as grating and filter changes. These digitization and sequence steps were later eliminated by the introduction of a high-performance instrument (Perkin-Elmer Model 180) which was designed with built-in digital electronics. At first, data were transferred from the spectrophotometer to magnetic tape for off-line processing. It was soon recognised that the spectroscopist prefers to work within the instrument laboratory and with the data processing preferably carried out in the same area, i.e. an on-line data system was required. This was confirmed by the increased interest generated by workers who were using Fourier Transform (FT) instruments. Here, the application of an on-line computer was essential for the production of a conventional absorption spectrum from the interferogram. One initial reaction was to consider connection of the instrument via a modem or direct data line to a time-shared computer. In many cases this was proposed where large companies and universities had central computing facilities. This approach was sometimes successful but often cost and inconvenience were important drawbacks. The main gain in the use of this style of system was for the execution of lengthy programs or for programs requiring a large storage capacity. Bearing the

spectroscopist's requirements in mind and with the availability of low-cost minicomputers, a full data processing package was developed for the Model 180 based on the Interdata 7-16 (later the 6-16) minicomputer. An instrument interface was also constructed — conforming to the RS232C/CCITV-24 specification — to allow direct coupling between instrument and computer. The RS232C specification was chosen since it was a well defined I/O specification used for data terminals and could be interfaced to most computers (even small programmable calculators). This provided an asynchronous serial link between instrument and computer. The software was written in an extended version of Fortran IV with a modular construction of linked sub-routines to enable easy adjustment and extension to the program. It was obvious from this early work that many new possibilities were now open to the spectroscopist and a new era in infrared spectroscopy was developing. To enable the benefits of data processing to be extended to routine analysis, a series of low-cost systems for computerized dispersive spectroscopy (c.d.s.) were developed based on cheaper instruments, e.g. Models 28X and 580. These systems are featured in the following discussions.

CURRENT HARDWARE FOR INFRARED DATA SYSTEMS

The c.d.s. data systems involve the use of a standard infrared instrument, equipped with the RS232C (V24) interface and connected to a standard minicomputer. This is an important factor since neither instrument nor computer are dependent on each other and can be used for independent operation; yet when combined, the computer processing facility greatly extends the performance of the instrument. Similarly, the configuration is not strictly limited to the use of a specific computer because the instrument can be readily interfaced to other modern computers that possess similar specifications to the Interdata 6-16. The instruments used in these systems can be placed into two categories: the 28X series of microprocessor-controlled spectrometer designed with the standard optical null system; the 580, computer-controlled spectrophotometer designed with a ratio-recording system. Essentially, the data processing of the systems based on these categories is identical in operation with one major difference, where the Model 580 is also controlled by the computer. A further gain with the latter system is also experienced in certain cases where an increased sensitivity is achieved with the ratio-recording system compared with the optical null instrument.

A schematic diagram of a typical c.d.s. layout is given in Figure 1. Here, data are transferred down a cable from the interface to the computer. The interface output is arranged in such a way that the computer "sees" the instrument as a standard I/O device, i.e. like a teletype. Once in the computer, the data can be processed via commands entered from the teletype console. Processed data can then be transmitted back down the data-link, through the interface and replotted in a standard format on pre-calibrated chart paper on the instrument recorder. In the case of the Model 580, the data-link also serves to transmit control instructions to enable the operator to perform automatic

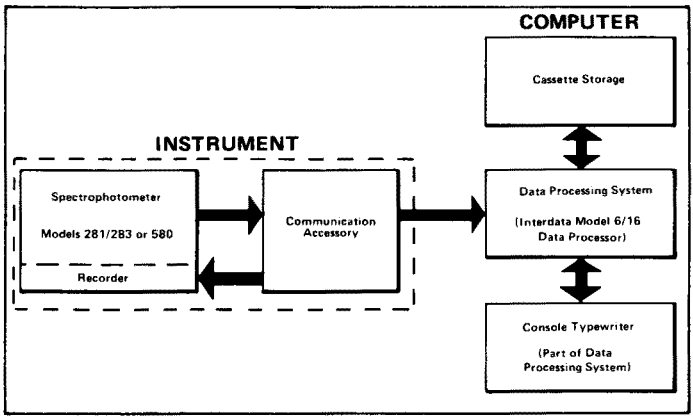


Fig. 1. A typical c.d.s. system.

setup of the instrument parameters (scan parameter, wavenumber, etc.) via the main teletype console or another peripheral device. The current system has the facility for program loading and bulk data storage on standard data cassettes. Future systems will feature "floppy" discs as an alternative (and preferred) medium for program transfer and data storage.

A microprocessor-controlled, 30 character per second type printer is also featured as the main system console. This style of printer is now preferred for speed and relative quietness when compared with the old-style all-mechanical teletype, although these can still be considered for use as the system console, if required. An alternative approach is to use a fast CRT display, which is absolutely silent and in many cases is cheaper than equivalent mechanical keyboard devices.

CURRENT SOFTWARE FOR INFRARED DATA SYSTEMS

In an attempt to make the main spectroscopic software very flexible and highly portable, i.e. easily converted for use with other data systems, the main program is structured with a series of sub-routines relating to specific functions within the software. Also, a major proportion of these routines is written in Fortran to simplify cross-translation for use with other computers. A prime consideration in the development of laboratory-based software is the interface of the program with the operator. Since many operators may have only a limited knowledge of computer operation, it is necessary to devise a simple operating language. Most of the commands used in the c.d.s. software (known as SPECT) are simple English words (or condensed forms of English words) that directly relate to the spectroscopic function being performed. At present approximately forty of these commands are available for data transfer, spectral manipulation, data storage and instrument control (in the case of the Model 580). Operation simply involves entry of the appropriate command modified,

if necessary, by a suitable argument string to perform the required function. For example, the command SCAN A,4000,600,1 prepares the system for a spectrum scan. The data will be transferred to an assigned area of computer memory, indicated by"A", and the spectrum is to be recorded from 4000 cm^{-1} to 600 cm^{-1} with the spectral information being sampled once every wavenumber (data interval = 1 cm^{-1}). A large protected area (9 K—16 bit word) of core memory is reserved for data collection, manipulation and temporary storage. For convenience with data manipulation, this is split into three working areas which by default are set to a capacity of 3000 data points. The size of each area is completely flexible and may be freely varied up to the maximum of 9000 points at the sacrifice of the other two areas. The command PLOT A will initiate a replot of stored and/or processed data from the assigned area "A" on the instrument recorder in calibration, unless alternative scaling parameters are set for the intensity or wavenumber scales (command called SCALE).

In the event of an operator error or an instrumental fault, an error is indicated in the form *ERROR* IN CMD: X; where CMD indicates the command in which the error occurs (e.g., SCAN, PLOT, SCALE, etc.) and X is a code number specific to the source of the error. Hence in this way rapid diagnosis of syntax and hardware faults is simply achieved.

APPLICATION OF DATA PROCESSING TO SPECTROSCOPIC PROBLEMS

Numerous applications of computers to spectral data processing have appeared in the literature during the past two years. For illustration, four examples of spectrum manipulation are discussed below.

Computer-calculated difference for the spectroscopic separation of drug mixtures

In this example the spectrum of a mixture of 80% procaine hydrochloride and 20% lidocaine hydrochloride is recorded by a Model 283-based system (CDS 3). The sample is prepared as a standard 13 mm KBr disc (Fig. 2a). A second spectrum of relatively pure procaine hydrochloride is recorded by a similar procedure. The computed difference spectrum (Fig. 2b) which relates to the remaining lidocaine hydrochloride, is generated by a scaled absorbance subtraction of the procaine hydrochloride spectrum from the mixture spectrum. Most of this computation is initiated by the use of a single command in the form DIFF A, B, ν_1, ν_2; the command DIFF (for DIFFerence) automatically converts the two spectra A (mixture) and B (pure compound) into the absorbance form. The reference or pure compound spectrum (B) is then multiplied by a factor (scaled) which equalizes the intensity of a common band (present in A and B) defined by the frequency limits ν_1 and ν_2 (chosen by the operator). After this multiplication, spectra A and B are subtracted and the result, still in the absorbance format, is placed into an area in the computer memory without any corruption of the original data. The absorbance format is retained

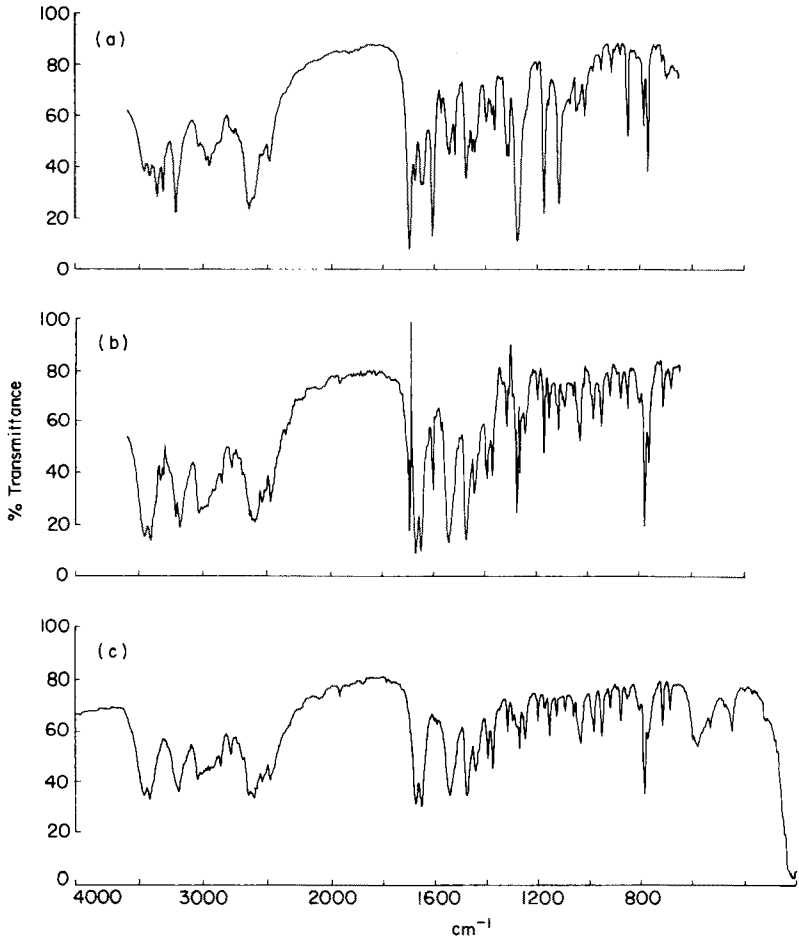


Fig. 2. Spectroscopic separation of a mixture. (a) Spectrum of 80% procaine · HCl—20% lidocaine · HCl mixture (KBr disc); (b) computer difference spectrum (mixture — procaine · HCl); (c) reference spectrum for lidocaine · HCl (KBr disc).

to enable subsequent scaling if necessary before returning to the conventional percent transmittance format. This is performed by the command ABST (ABSorbance → Transmittance). Hence the difference data in Fig. 2(b) are produced by two simple commands. A good correlation is obtained between this spectrum and a reference spectrum of the lidocaine hydrochloride (Fig. 2c).

In this context, the technique is valuable for both the analysis of mixed drug formulations and in the detection and identification of impurities (or metabolites) in pharmaceutical products. It is also worth noting that this exercise with compressed halide discs would be practically impossible without computer data handling. Further practical applications of computer difference are reported separately [1].

Application of signal averaging in sub-micro analysis

In this experiment, the spectrum of 10 ng of amobarbital was produced from a 0.5-mm microdisc situated in an 8X beam condenser, with the Model 283-CDS 3 instrument. It is evident from Fig. 3(a) that there is very little indication of the amobarbital when the sample spectrum is compared to the background.

Instrumental noise could have a significant contribution in the final spectrum if the data are extracted by direct computer subtraction of the background from the sample spectrum. The significance of the spectral data can be improved by the signal averaging routine which is available in the c.d.s. software. The result of background subtraction following 16 averaged scans of sample and background with 12X ordinate expansion is shown in Fig. 3(b). All the major absorptions of the drug are easily observed even at the low level of 10 ng.

This facility for providing data enhancement by both scale expansion and signal averaging with the aid of the computer has enabled infrared spectroscopy to be used in areas traditionally served by gas chromatography. A more comprehensive coverage of this technique has been given [2].

Application of signal averaging and digital smoothing to microanalysis by surface reflectance

Figure 4 (spectrum A) shows the quality of data obtained by direct reflectance from a 2-mm sample of a painted surface. The maximum apparent transmittance level, at best, was only 0.4%T. Normally, under these conditions, data would be subject to errors and distortions caused by lack of instrumental sensitivity and response near 0%T. In this case, however, an advantage is gained by the use of a ratio-recording system with the Model 580 (CDS 2). First attempts to expand spectrum A directly resulted in data which did not correlate with a control reference reflectance spectrum of the sample (Fig. 5, spectrum B).

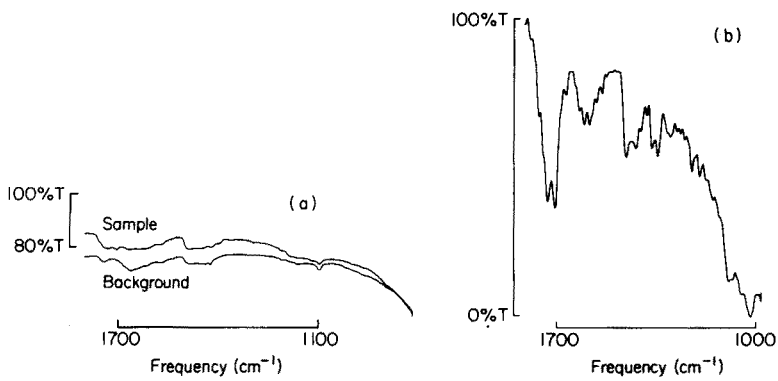


Fig. 3. Signal averaging. (a) The upper spectrum is for 10 ng of amobarbital while the lower spectrum is the background; (b) spectrum for 10 ng of amobarbital obtained from 16 scans averaged with 12X ordinate expansion.

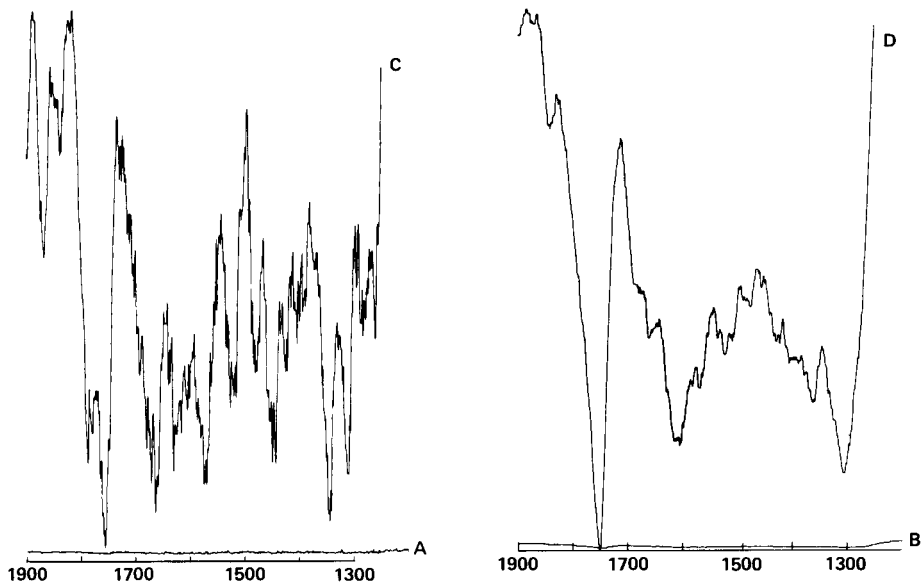


Fig. 4. Signal averaging for microreflectance measurements on a 2-mm paint sample. (A) Normal reflectance spectrum; (B) reflectance spectrum with minor attenuation and signal averaging; (C/D) spectra A/B with auto-expansion and 36-point smooth.

This happened because there is a major contribution from instrumental noise in the spectrum. The result of signal averaging on spectrum A is seen in spectra B and D (expanded form of B). There is a good correlation between this final spectrum, also illustrated in Fig. 5(A), and the standard reflectance spectrum of the sample (Fig. 5B). It should be noted that a further software routine — digital smoothing — is used to improve the appearance of the data in spectra C/D (Fig. 4) by the removal of some of the high-frequency noise components.

Application of computer control in automated analysis

The c.d.s. software permits the user to "stack" commands by the generation of an instruction input file prepared on paper tape or magnetic cassette (or in the future floppy disc). The file consists of a simple sequential list of the normal program commands such as SCAN, PLOT, DIFF, etc., set down in the sequence required for any specific analysis. Under normal operating conditions, the operator can first establish the optimum scan and processing parameters required for the analysis and then generate the ASCII file for future analyses under the same conditions. This approach can be extended by the use of an autosampler accessory which can also be controlled from the computer on the Model 580 system (CDS 2). In this way either related or unrelated analyses can be performed automatically without intervention of the operator.

To illustrate the point, Fig. 6 gives a few results produced from an automated run set up for the analysis of styrene—butadiene copolymers. Films

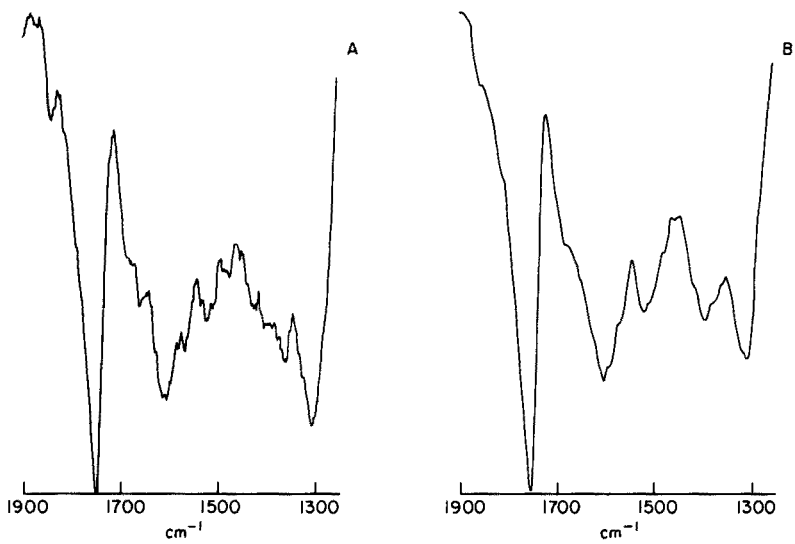


Fig. 5. Microreflectance measurements. (A) Spectrum for a 2-mm paint sample; (B) standard spectrum of a full-size sample.

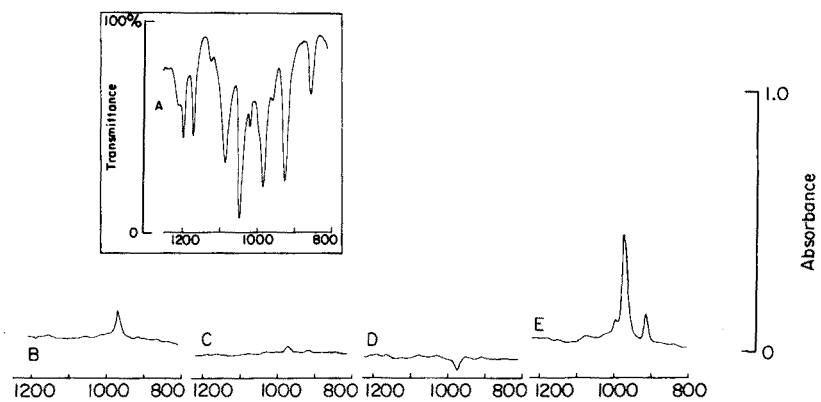


Fig. 6. Analysis of styrene—butadiene copolymers. (A) Reference sample; (B—E) samples (differences). Scan range, 1250—800 cm^{-1}; band scaling, 1029 cm^{-1}.

were examined from a series of production batches of the polymer. Each sample was analysed in turn by recording the spectrum and comparing this by the auto-scaled difference with the spectrum of a production control standard. The band selected for auto-scaling (1029 cm^{-1}) is from the styrene part of the polymer and on subtraction removes the spectral components relating to styrene plus associated butadiene in the control sample. Any residue left after subtraction reflects variations in butadiene content, which is an important factor in the production control of the final product. Spectrum

C (Fig. 6) indicates that this sample is very close in butadiene content to the control sample (spectrum A). Sample B contains more butadiene but is at a level that can be tolerated whereas sample E is far in excess; sample D contains less butadiene and is just within specification. These four samples form part of the series that were analysed fully automatically by the following command sequence:

SCAN A,1250,800,1	sets scan ranges
SCAN B,1250,800,1	and data increment
MODE 2	sets spectroscopic parameters
RUN B	initiates scan of control sample
SAMPLE	changes sample in autosampler
RUN A	initiates scan for first sample
DIFF A,B,C,1045,995	calculates difference,. autoscale on 1029 cm^{-1} band
PLOT C	plots difference
NAME C,SPLEØ1	gives serial number to spectrum
SAVE C, 1	stores spectrum on cassette
SAMPLE	
RUN A	continues sequence as above
etc., etc.	

This is an example of a related series for analysis. A similar routine could be adopted for an unrelated series merely by changing the scan limits and the processing commands between each sample change. This type of analysis not only reduces the degree of operator interaction but also enables instruments with a heavy work load to be utilized over a 24 h period.

Conclusion

The development of a commercial system for spectral processing of infrared data has been highlighted. With careful attention to sample preparation and manipulation of the infrared data with the standard applications software, a wide variety of new practical applications of infrared spectroscopy can be considered. Furthermore, the full intrinsic capabilities of an instrument can be readily achieved and extended beyond the normal performance levels by the use of simple computer routines.

REFERENCES

1 J. P. Coates, Anal. Chim. Acta, 103 (1978) 000.
2 J. P. Coates, Int. Lab. (1977)

Analytica Chimica Acta, 103 (1978) 313—321
Computer Techniques and Optimization

NUMERICAL TAXONOMY AND INFORMATION THEORY APPLIED TO FEATURE SELECTION FROM FILED INFRARED SPECTRA FOR AUTOMATED INTERPRETATION

F. H. HEITE, P. F. DUPUIS[‡], H. A. VAN 'T KLOOSTER* AND A. DIJKSTRA

Analytisch Chemisch Laboratorium, Rijksuniversiteit Utrecht, Croesestraat 77A, Utrecht (The Netherlands)

(Received 3rd May 1978)

SUMMARY

A method is described for the selection of features from infrared spectra, aimed at computer-aided interpretation by retrieval of coded spectra. The coding procedure is similar to that of the ASTM Infrared Spectral Index, involving 140 binary-coded wavelength intervals (peak positions) of 0.1 μm for each spectrum (Wyandotte code). In addition to this procedure, windows of 0.3 μm and 0.5 μm are used. For a given set of spectra, the peak positions are grouped by means of numerical taxonomy; the correlation coefficient is used as a criterion. Information contents of all peak positions are calculated with Shannon's formula. One peak position is selected from each group, viz. the position having the highest information content. The selection obtained in this way is composed of peak positions that are weakly correlated yet yield much information. The spectra are then coded again, taking account of the selected peak positions only. In order to evaluate the selection, the number of spectra still differing from all other spectra in the set is determined by comparing all reduced spectra with each other. For a file of 395 spectra (hydrocarbons, alcohols, ethers, carbonyl compounds) 99.0% of the spectra are unique when 27 selected features are used. For a file of 5100 spectra (of a wide variety of compounds, taken from the ASTM Infrared Spectral Index) 97.7% of these spectra are unique when only 40 out of all 140 peak positions are used.

The investigations reported in this paper are part of a project which aims at the development and evaluation of strategies for the optimization and selection of analytical procedures, including feature selection, coding, and interpretation of analytical data, with the application of chemometric methods [1—5]. Numerical taxonomy, as defined by Sneath and Sokal [6], was introduced into analytical chemistry by Massart and de Clercq [7] in an application to the classification of thin-layer chromatographic systems. Information theory and numerical taxonomy were applied [2] to the selection of stationary phases for gas—liquid chromatography. Information theory was applied [4] to an evaluation of the ASTM Infrared Spectral Index for retrieval purposes.

A main aspect in computer handling of infrared spectra is how to reduce and code the spectra efficiently and effectively, to achieve optimum

[‡]Present address: Dow Chemical (Nederland) B.V. Terneuzen (Netherlands)

performance in data reduction and information retention. In the ideal case unambiguous identification of compounds by retrieval of their coded infrared spectra requires, as an essential condition, that the reference spectra of the file being used are all uniquely coded. To this effect a feature selection procedure must deal with correlations between the features present in the original data.

This paper describes the application of numerical taxonomy to the grouping of spectral features (binary-coded peak positions), with the correlation coefficient taken as a criterion of similarity. The information content of the features, calculated from Shannon's formula [8], is used as a selection criterion.

Coding the spectra

The binary code used is based on the division of the wavelength range 2.0—15.99 μm ($\hat{=}$ 5000—625 cm^{-1}) into 140 intervals (peak positions) of 0.1 μm each. If a spectrum shows one or more peaks in a certain interval above a preset intensity threshold, the code "1" is assigned to this position. If there is no peak above the threshold, the code is "0". This is the Wyandotte ASTM code or 0/1 code [9]. The intensity of a peak is defined as the difference between the transmittance level of the base line of the spectrum and the transmittance of the top of the peak (the transmittance ranging from 0 to 100%). Errors introduced in recording and coding a spectrum may cause a peak to appear at a position next to the "true" position. To obtain a "1" at the "true" interval also, a window can be used. When a window of width 0.3 μm is applied, the positions on the left- and right-hand sides of the interval that shows a peak are also given the code "1". This is called the 0.3-μm code. The procedure can be extended to two positions on the left and two on the right-hand side of a peak position: this gives the 0.5-μm code. A schematic presentation is given in Fig. 1.

NUMERICAL TAXONOMY

The grouping method called numerical taxonomy is defined as the grouping by numerical methods of operational taxonomic units (OTU's) on the basis

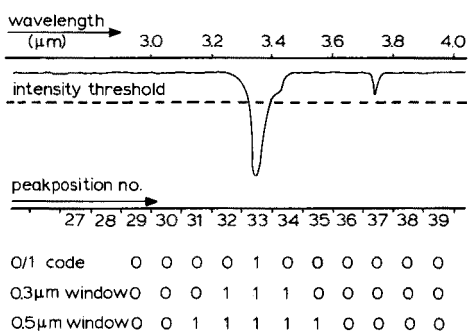


Fig. 1. Part of an infrared spectrum (schematic), binary-coded in three ways.

of their character states [6]. Spectra can be arranged as OTU's on the basis of the peaks they contain: the most similar spectra form a close group, whereas dissimilar spectra fall into different groups. This paper deals with the classification of peak positions on the basis of their presence in a given set of spectra. The correlation coefficient was chosen as a criterion for the "resemblance" of the peak positions; other criteria such as taxonomic or Euclidean distance could also be chosen. Because there are 140 peak positions, a symmetrical 140 × 140 matrix was employed.

The classification is performed in the following way: the highest correlation coefficient $r(i, j)$ is selected: i and j are the most similar positions and are considered to form one group i'. Then, the correlation coefficients between this new group and all other positions are calculated (this can be done in several ways; see below). In the matrix, row and column j can be erased; row and column i are replaced by the newly calculated values for group i'. Thus, the dimension of the matrix is reduced by one. In the reduced matrix, the highest correlation coefficient is selected again and this procedure is repeated until (after 139 steps) all 140 peak positions are classified into one non-overlapping hierarchical system of groups and subgroups.

In this investigation, two methods were used for calculating the correlations between a newly formed group i' and all other groups k. In the "Weighted Pair Group Arithmetic" (WPGMA) each cluster is given equal weight in determining the linkage levels, regardless of how many peak positions are contained in the groups:

$$r(i', k) = [r(i, k) + r(j, k)]/2 \tag{1}$$

In the "Unweighted Pair Group Arithmetic" (UPGMA) each group is weighted in proportion to the number of peak positions it contains:

$$r(i', k) = [n_i r(i, k) + n_j r(j, k)]/[n_i + n_j] \tag{2}$$

where $r(i, k)$ is the correlation coefficient between groups i and k, and n_i is the number of positions contained in group i.

The cluster process can be represented graphically in a dendrogram (tree); an example is given in Fig. 2. The OTU's are given along the x-axis; the y-axis shows the correlation levels of the mergers between the OTU's.

INFORMATION CONTENTS OF PEAK POSITIONS

With Shannon's formula [8], the information content I_i of each of the 140 peak positions can be calculated for a given set of spectra:

$$I_i = p_i \,.\, ld(p_i) - (1 - p_i)\, ld(1 - p_i) \tag{3}$$

where p_i is the number of coded spectra having a "1" at position i, divided by the total number of spectra in the set, and $ld = \log_2$.

In calculating the information contents of two or more positions, the algebraic sum of these values must be corrected for the correlations between

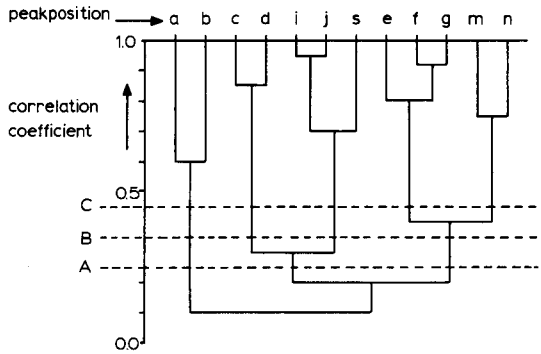


Fig. 2. Dendrogram of peak positions (schematic), classified on the basis of correlation coefficients. Illustration of feature selection by considering cross-sections (lines A, B and C).

these peak positions. When positions l through n are taken together, their net, total amount of information, I_{nt}, is given [4] by:

$$I_{nt} = \sum_{i=1}^{n} I_i + \tfrac{1}{2}\, ld \begin{vmatrix} 1 & r(1,2) & \cdots & r(1,n) \\ r(2,1) & 1 & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ r(n,1) & \cdot & \cdots & 1 \end{vmatrix} \qquad (4)$$

where $r(i,j)$ is the correlation coefficient of peak positions i and j. Instead of calculating this determinant, a simpler solution was used. Let r_m be the largest correlation present in the determinant. Now, all values $r(i,j)$ can be replaced by r_m. The determinant then equals $(1 - r_m)^{n-1} \cdot [(n-1) \cdot r_m + 1]$, which gives a large saving in computer time. The correction values obtained in this way are, however, somewhat larger than the real values, so that the net total amounts of information become slightly more pessimistic.

METHOD OF SELECTION

A cross-section of the dendrogram (see Fig. 2, for instance at level A) provides a certain number of relatively weakly correlated groups, each group consisting of one or more relatively highly correlated OTU's (peak positions). Now from each group the position having the highest information content is selected. If the resulting number of positions is too small, a larger number of groups can be obtained by shifting the cross-section level in the dendrogram (lines B and C). Compared with the previous selection, these new selections contain more peak positions and thus more bits of information. Conversely, if the number of groups is too large, the level must be shifted downwards.

The selection is stopped when a certain preset number of positions is reached, or when a certain preset net total amount of information is obtained.

In the results given below, selection was terminated when a net amount of $ld(N)$ bits of information was obtained, where N denotes the number of spectra contained in the file. When the corresponding N compounds are considered to be the population of all compounds, and when all spectra are without errors, an amount of $ld(N)$ bits allows discrimination of all spectra from each other.

In order to prevent a selection from containing a large number of peak positions with small information contents, positions having less than a certain amount of information are excluded. A possible information threshold is the average information of all 140 peak positions.

Evaluation of a selection

A selection made for a given set of spectra is evaluated with respect to the number of spectra that still differ from all other spectra, when reduced to only the selected features. (The other features yield little information or are highly correlated to the selected features.) Thus, all spectra in the set are coded again, only the selected peak positions being taken into account. All the reduced spectra obtained in this way are compared with each other and the following two parameters are calculated: (a) the percentage of uniquely coded spectra; and (b) the so-called degeneration of the spectra, averaged over the total number of spectra in the set.

A spectrum is unique if its specific combination of ones and zeros (as for the selected positions) differs from every other coded spectrum within the set. A reduced spectrum is degenerated x-fold if $(x - 1)$ other spectra show the same combination of ones and zeros; a 1-fold degenerated spectrum therefore is unique within the set.

Methods

Calculations were performed on the CDC-Cyber 73/26 computer of the Academic Computer Centre of the State University of Utrecht; programs were written in Fortran IV. For printing dendrograms on the line printer, a program by Davis [10] was used. For the program performing the numerical taxonomy, some routines were derived from a publication by Anderberg [11].

The data sets used are listed in Table 1.

TABLE 1

Data sets used

Ref.[a]	No. of spectra	Compounds
A5100	5100	Miscellaneous
B395	395	Hydrocarbons, alcohols, ethers, aldehydes/ketones
C164	164	Hydrocarbons

[a] A: ASTM infrared spectral index. B and C: collection of IR—Raman Department, Utrecht University.

RESULTS AND DISCUSSION

Windows

Table 2 shows the results obtained for data sets A5100. By the nature of the 0.3 and 0.5-μm codes, the number of ones in the coded spectra is increased; consequently, the information contents of the peak positions are larger than when the 0/1 code is used. For this reason fewer peak positions are required when a window is involved. As for the number of unique reduced spectra, the 0/1 code gives the best results, however, with 25% more features than the other two codes. The degree of degeneration is decreased strongly by the use of a window. On the basis of these results, the 0.3 or 0.5-μm code should be preferred. However, if errors in the sets and correlations between the 140 peak positions are considered (according to Dupuis and Dijkstra [4]), the total information content of the set with the 0.3-μm window is larger than for the other sets. Eventually preference was given to the 0.3-μm window.

Intensity threshold

Table 3 gives the results for the sets B395 and C164, when different intensity thresholds, viz. 5, 10 and 50%, were used. The calculations were carried out with both weighted and unweighted clustering. When a threshold of 50% is used, large parts of the spectral region are empty: 50 peak positions for set

TABLE 2

Results for data sets A5100
(Intensity threshold, 5%; cluster procedure, WPGMA)

Window used (μm)	—	0.3	0.5
Information threshold I_{th} (bits)	0.48	0.72	0.76
No. of selected features	19	14	14
Percentage of unique spectra	45	38	39
Average degeneration	13.2	3.6	3.3

TABLE 3

Results for data sets B395 and C164
(0/1 code — no window)

	Data set B395					Data set C164			
Intensity threshold (%)	5	10	50	5	10	5	10	5	10
Clustering procedure[a]	W	W	W	U	U	W	W	U	U
Information threshold (bits)	0.48	0.42	0.22	0.48	0.42	0.43	0.34	0.43	0.
No. of selected features	11	11	19	11	12	8	11	10	14
Percentage unique spectra	43	36	44	53	47	34	38	44	52
Average degeneration	2.9	4.5	9.6	2.3	7.7	3.2	3.3	3.2	7.

[a]W = weighted pair group arithmetic. U = unweighted pair group arithmetic.

C164 and 23 for set B395. Besides, the number of peaks at the other positions is very small, giving them low information contents. As a consequence the number of features to be selected is increased, and also their in-between correlations. In some cases no selection could be made: the amount of information to be subtracted, because of correlations, became so large that the preset quantity of $ld(N)$ bits could not be reached. As can be seen from Table 3, for set B395 a threshold of 5% gives the best results. For set C164, a value of 5% gives the best values for the degeneration, whereas a threshold of 10% gives the highest number of unique spectra. It appears that an intensity threshold of 5% should be preferred.

Weighted vs. unweighted clustering

From Table 3, it can be derived that the unweighted method (UPGMA) gives the best results. This agrees with a statement by Sneath and Sokal [6] that, for clustering binary coded taxonomic units, UPGMA gives better results than WPGMA.

Information contents of features and combinations of features

In obtaining the results shown in Tables 2 and 3, the average information content of all 140 peak positions was used as an information threshold for the features to be selected. These average values range from 0.22 to 0.76 for the different data sets. An alternative is to use a fixed information threshold, to be chosen more or less arbitrarily, but in any case set rather high to allow the selection of features with relatively high information contents only. In the investigations discussed in the following paragraphs, an information threshold of 0.88 bit is used; this implies that only positions present in at least 30% and at the most 70% of the spectra are admitted to the selection (see eqn. 3).

Although essentially a total amount of $ld(N)$ bit should be enough to discriminate N spectra, this was not the case as can be concluded from the results given above. Apparently, calculation of the information content of n peak positions by eqn. (4) is not correct. In fact, for n selected features, eqn. (4) is valid only if N is much greater than 2^n. This means, unfortunately, that eqn. (4) is inadequate if the aim is to achieve 100% uniquely coded spectra. Work to optimize the model for calculation of information contents is in progress.

Percentage of unique spectra vs. number of selected features

In order to determine how many (or rather, how few) selected features are required to obtain an acceptable percentage of uniquely coded spectra, calculations were carried out involving data sets A5100 and B395. In these calculations, the conclusions drawn above were applied, i.e. coding according to the 0.3-μm method, intensity threshold 5%, information threshold 0.88 bit, and unweighted clustering. The results for data set B395 are given in Fig. 3. It appears that 99% of the spectra are uniquely coded by 27 selected features.

With 35 peak positions, only 3 out of 395 spectra are still identical (two of which were already identical before reduction). As this result was judged to be sufficient, feature selection was stopped at this number of 35 peak positions.

For data set A5100 the results given in Fig. 4 show that 97.7% of the 5100 spectra are uniquely coded with 40 selected peak positions, at which the selection was stopped, in order to cut down computer time. The degeneration, averaged over all 5100 spectra, amounted to 1.07. This means that a forward search with an unknown spectrum, on average, will yield 1.07 spectrum with a match of 100%, provided that a reference spectrum of the "unknown" compound is present in the file and no account is taken of the errors introduced in recording and coding of unknown and reference spectra. (Previous work [4] has revealed how dramatic the effects of errors and correlations are with respect to the net information content of infrared spectra files and their use for retrieval purposes.)

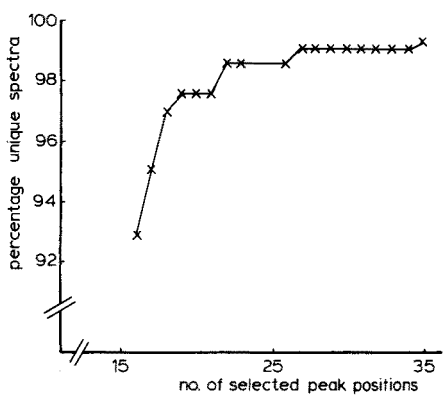
To some extent, a correction was made for such errors by using a window in coding the peak positions. Nevertheless, the importance of a well-described, standardized recording procedure must be emphasized.

Infrared spectrometric relevance of selected features

In Fig. 5 the positions of the selected features on the wavelength (or wavenumber) scale are indicated. A vast majority of the selected features are in the fingerprint area. This may be important, considering the possibility of combining a chemometric treatment with the spectrometric ab initio interpretation of infrared spectra. The results of such different approaches could well be complementary.

Conclusions

A considerable part of infrared spectral data, contained in files and binary-coded according to the ASTM-Wyandotte method, appears to be redundant. Aiming at computer-aided processing and interpretation of infrared spectra, the supposed need for data reduction may be argued at present in view of



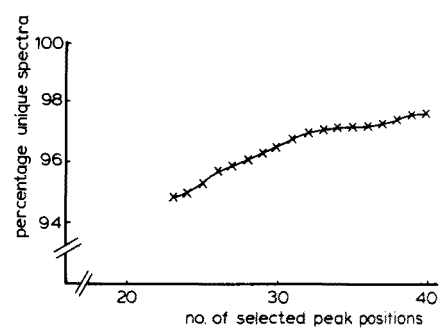


Fig. 3. Plot of the percentage uniquely coded spectra vs. number of selected peak positions for data set B395.

Fig. 4. Plot of the percentage uniquely coded spectra vs. number of selected peak positions for data set A5100.

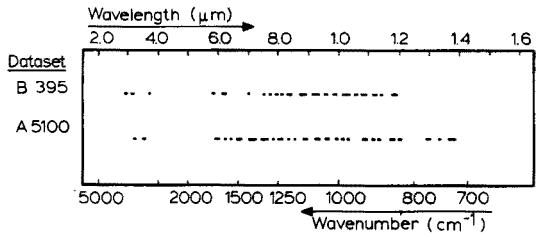


Fig. 5. Positions of selected features on the wavelength and wavenumber scales for data sets A5100 and B395.

the substantially decreasing costs of computer hardware. If, however, data reduction is considered to be necessary (e.g. for reasons of software development), numerical taxonomy and information theory have been shown to be useful in feature selection.

The authors are indebted to Dr. J. H. van der Maas for kindly supplying some data sets, to Professor D. L. Massart for providing a program for numerical taxonomy which was used in preliminary calculations, and to Ing. J. Schouten and Drs. P. Cleij for valuable discussions

REFERENCES

1 P. F. Dupuis and A. Dijkstra, Anal. Chem., 47 (1975) 379.
2 A. Eskes, P. F. Dupuis, A. Dijkstra, H. de Clercq and D. L. Massart, Anal. Chem., 47 (1975) 2168.
3 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
4 P. F. Dupuis and A. Dijkstra, Fresenius Z. Anal. Chem., 290 (1978) 357.
5 P. F. Dupuis, A. Dijkstra and J. H. van der Maas, Fresenius Z. Anal. Chem., 291 (1978) 27.
6 P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy, Freeman, San Francisco, 1973.
7 D. L. Massart and H. de Clercq, Anal. Chem., 46 (1974) 1988.
8 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, The University of Illinois Press, Urbana, Ill., 1949.
9 Codes and Instructions for Wyandotte—ASTM Punched Cards, American Society for Testing and Materials, Philadelphia, Pa, 1964.
10 J. C. Davis, Statistics and Data Analysis in Geology, Wiley, New York, 1973.
11 M. R. Anderberg, Cluster Analysis for Applications, Academic Press, New York, 1973.

Analytica Chimica Acta, 103 (1978) 323—338
Computer Techniques and Optimization

INDUSTRIAL APPLICATIONS OF COMPUTERIZED DISPERSIVE INFRARED SPECTROSCOPY FOR ANALYSIS IN SOLUTION

J. P. COATES
Perkin-Elmer Limited, Beaconsfield, Buckinghamshire (Gt. Britain)
(Received 31st May 1978)

SUMMARY

Today, with the relatively low cost of data processing equipment, it is possible to interface a dedicated minicomputer to a computer-compatible scientific instrument to give a total system that is ideal for routine operation. This article describes some basic chemical applications of a commercial system for computerized infrared spectroscopy that are relevant to analytical problems encountered in industrial laboratories. In the initial stages of the article some examples are given of the different experimental procedures that are available to handle infrared data from solutions. The remaining sections illustrate the direct application of data handling to specific applications in both organic and aqueous media.

Nowadays, many infrared analyses are performed directly on the sample "as received" without previous dilution in a solvent. This contrasts with other spectroscopic techniques, such as u.v.-visible, atomic absorption, fluorescence, n.m.r., etc., where samples are usually prepared as solutions in suitable solvents. In many cases the dilution step is omitted for convenience, since there is not an "ideal" infrared solvent that has complete transparency throughout the normal infrared region. Also, conventional infrared instrumentation often lacks the sensitivity required to handle dilute solutions. There are, however, numerous experiments that require the analyst to produce infrared spectroscopic data from solution: for example, analysis of samples submitted as solutions; quality control of products manufactured as solutions; studies on compounds unstable in the isolated state; critical analysis of infrared band shape, intensity and/or position; quantitative analysis, where careful concentration control is necessary; studies on solvent and/or solute interactions. Frequently the results from such studies are disappointing owing to difficulties associated with interference from solvent absorption bands and the inability to carry out further processing of the data from the solute. This paper indicates how data processing can be utilized to improve matters with the aid of practical examples, and illustrates how the technique may be adapted to many industrial applications. All experiments discussed in the text were performed on a commercial system based on the Perkin-Elmer CDS-2 infrared spectrophotometer (Model 580) connected to a minicomputer (Perkin-Elmer - Interdata 6/16).

For convenience, the discussion is divided into two sections. The first deals with various aspects of infrared analysis in simple organic solutions. The ideas from this section are extended in the second part to produce data from typical

samples encountered during routine industrial analysis. The potential of this work is far-reaching, and complex studies have evolved from many of the preliminary experiments.

INFRARED ANALYSIS OF SIMPLE ORGANIC SOLUTIONS

Solvent cancellation

The double-beam configuration of modern infrared instruments enables experiments, e.g. solvent cancellation and "differential" infrared spectroscopy, to be carried out. Here, the sample, in a standard cell, is compared directly with the reference material (often the solvent) in either a matched or variable pathlength cell. Results obtained by this procedure can vary considerably, since they often depend on the skill of the operator, the sensitivity of the instrument and the success of matching the sample and reference. A standard experiment is the removal of the background spectrum of a solvent such as carbon tetrachloride or chloroform with the aid of a variable pathlength cell. The exact cancellation by the above method can be tedious and inconvenient but, in contrast, the exercise becomes trivial with the aid of computer processi as illustrated in Fig. 1. In this case the spectra of two 5% solutions of benzyl acetate in carbon tetrachloride (A) and chloroform (B) are recorded from the same cell (50 μm, KBr) and the data are stored. Spectra of the pure solvents are similarly recorded and subsequently used for solvent cancellations with the autoscaled difference. The bands selected for solvent elimination are:

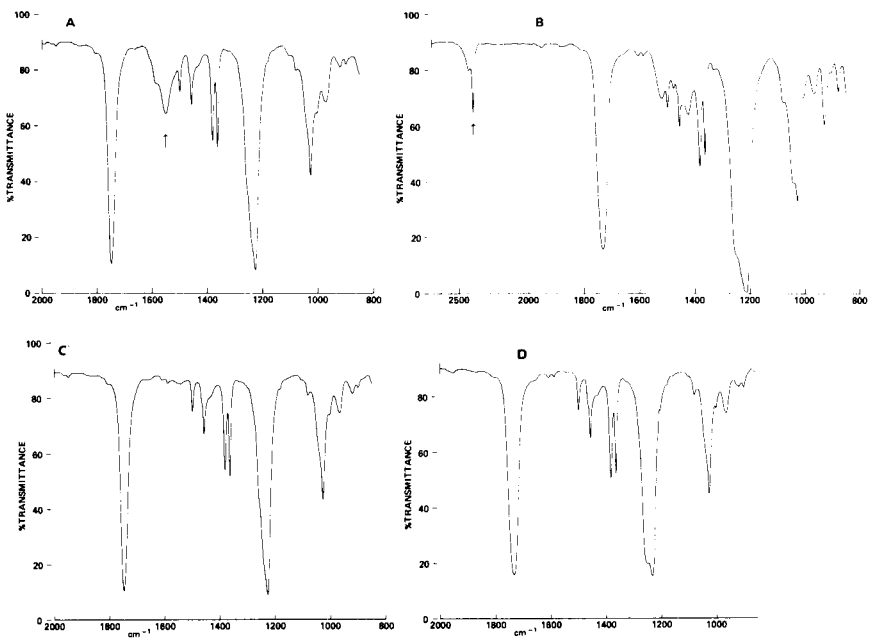


Fig. 1. Solvent cancellation by autoscaled computer difference. (A) 5% Benzyl acetate in carbon tetrachloride; (B) 5% benzyl acetate in chloroform; (C/D) spectra A and B after solvent subtraction. Arrows mark the bands used for autoscale.

1550 cm^{-1} (carbon tetrachloride) and 2400 cm^{-1} (chloroform). Solvent-free solute spectra are obtained for both solutions (spectra C/D), and the only practical limitation for this type of manipulation are the spectroscopic restrictions imposed by the areas of solvent cut-off, i.e. between 800 cm^{-1} and 700 cm^{-1} for carbon tetrachloride and chloroform, which have been excluded for this reason from Fig. 1.

Three advantages are gained from processing solution data in this way: (i) speed and convenience, (ii) elimination of any cell effects because the same cell is used throughout, and (iii) the possibility of carrying out subsequent data processing on the solvent-cancelled spectra.

Computer simulation of spectral band shifts

The occurrence of band shift in the infrared spectrum can be interpreted in many ways depending on various physical and chemical factors that can influence the sample; two examples are discussed later where these shifts are induced by interactions in solution. Often the shifts involved only correspond to one or two wavenumbers and are sometimes difficult to observe directly. The use of computer difference is valuable for observing any changes in band position since only the modified data appear in the difference spectrum usually in the form of a first derivative. Formulae are available for calculating the degree of shift from the derivative data [1], but these involve measurements from both the original and the derivatised spectrum and also require judgement of band shape. SPECT 580 software permits direct simulation of the shift for any band under investigation so long as the band is defined by an adequate number of points to correlate with the shift.

Two examples of computer simulation are produced in Fig. 2. The results illustrate the effect of band width and symmetry on the resultant derivative data. In this way, shifts may be estimated without prejudgement or measurements based on previously recorded data.

Direct comparison of solvent-induced shifts

A point mentioned above that relates to the use of data processing after solvent cancellation can be appreciated from the spectrum shown in Fig. 3, which is produced from the difference calculation on the two solute spectra from Fig. 1 (C/D). In this experiment the influence of the solvent on the solute spectrum can be observed directly in terms of the derivative data. Only the spectrum modifications are observed in the difference spectrum, and the result may be related to the relative effects of the solvents on the individual functional groups within the molecule. As expected, the result indicates that the greatest shifts are shown by the ester bands. In the event that no solvent interaction is experienced, the difference approximates to a straight line through zero absorbance in the relevant portions of the spectrum.

Studies of this nature give valuable information for many theoretical and practical infrared experiments and help to explain numerous unexpected results that occur when samples are examined in liquid media.

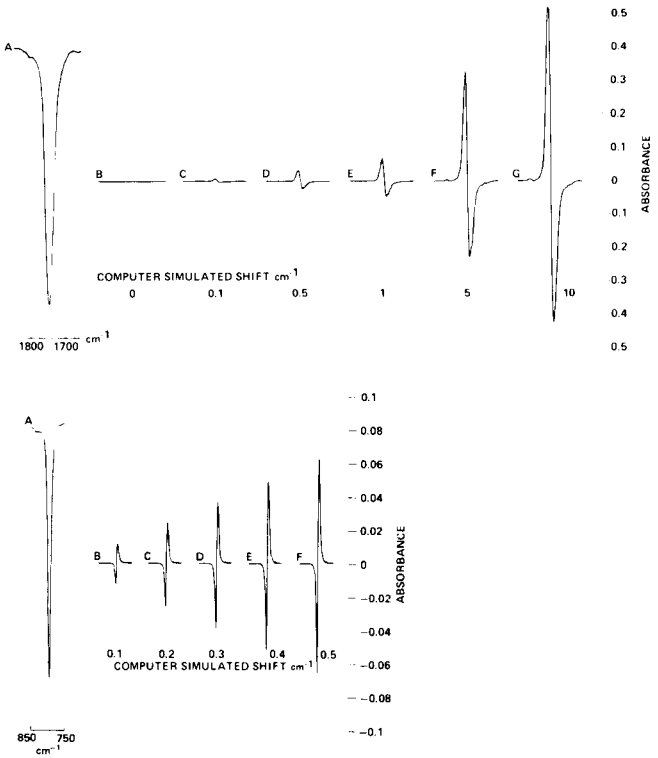


Fig. 2. Computer simulation of band shift. The upper part of the Figure shows: (A) original band (carbonyl absorption) for benzyl acetate in CCl_4; (B)–(G) computer difference before and after shift. The lower part of the Figure shows: (A) original band (ring C—H absorption for *p*-xylene; (B)–(F) computer difference after shift.

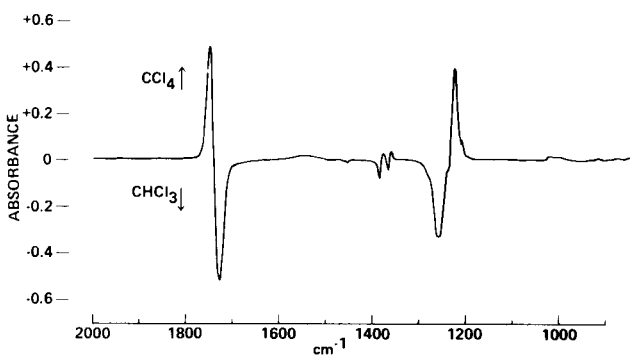


Fig. 3. Solvent-induced shifts for benzyl acetate; direct comparison of spectra C/D (Fig. 1) with automatic scaling on bands between 1400 and 1350 cm^{-1}.

Mixed solvent systems

Many of the points discussed previously apply to the study of mixed solvent systems. Further applications are considered for (a) a binary solvent mixture and (b) a commercial solvent mixture (xylol).

Spectroscopic separation of a solvent mixture. On occasions, solutions and solvent mixtures containing two or more components are presented for analysis where pure standards for comparison are unobtainable. Under these conditions it is generally necessary to adopt a separation technique, such as gas or liquid chromatography, to isolate the individual components for the production of the spectra of the pure materials. This is not always convenient because: complete separation may be tedious; components may be inseparable with conventional techniques; isolated components may be unstable in the isolated state; components may exist as a tautomeric or equilibrium mixture. Computer difference spectroscopy can be utilised here without the need for prior separation, so long as it is possible to vary the relative composition of the mixture. The last point is important because the exercise is simply a mathematical manipulation of simultaneous equations where the various coefficients are cancelled by successive autoscaled difference calculations. In principle, variations in composition are achieved by simple enrichment processes, e.g. azeotroping, partial adsorption on solid substrates, partial partition between an immiscible liquid interface, quenching a reaction (in the case of unstable species) or modification of an equilibrium with suitable catalysts.

The technique is demonstrated in Fig. 4 for two simple binary mixtures containing benzene and hexadeuterobenzene (A/B). Spectrum C of benzene is generated from spectra A and B by cancellation of the deuterated benzene

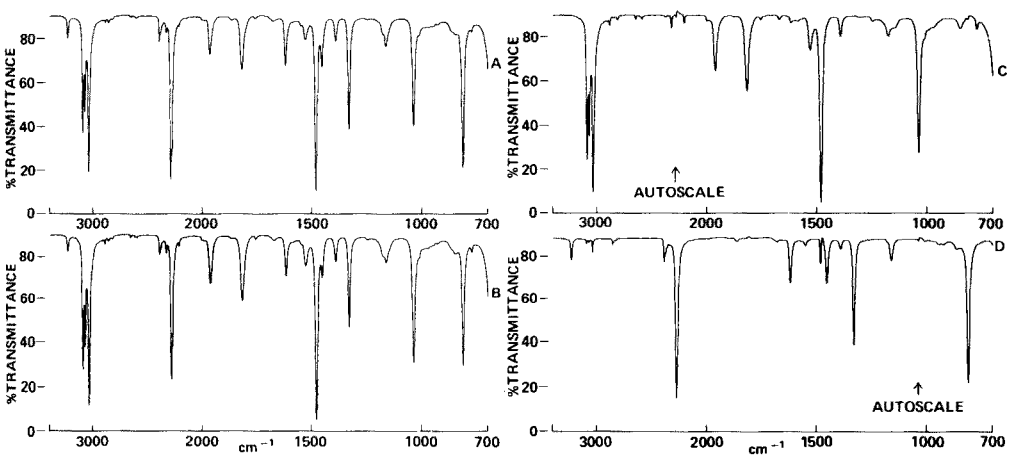


Fig. 4. Spectroscopic separation of a solvent mixture. (A) 40%/60% benzene—hexadeuterobenzene; (B) 60%/40% benzene—hexadeuterobenzene; (C) benzene spectrum generated from B—A with autoscale on 2280 cm^{-1}; (D) hexadeuterobenzene spectrum generated from A—B with autoscale on 1038 cm^{-1}.

based on autoscaling at the 2280 cm^{-1} band and subsequent subtraction of the two spectra. Similarly, the spectrum (D) of the deuterated species is obtained by subtraction of the spectra after autoscaling on the 1035 cm^{-1} benzene band.

Mutual interaction in solvent mixtures. The above procedure is also valuable when interactions occur between solvents in admixture. Theoretically, as soon as any mixture is formed, there is a degree of mutual interaction between components, even if there is a close chemical resemblance. Often attempts to carry out simple spectrum subtractions between pure compounds and composite mixtures fail to give complete cancellation for this reason. A typical mixture that exhibits this effect is commercial xylol which is mainly composed of the three xylene isomers with a small quantity of toluene. The subtraction of the spectrum of any one of these components from the xylol spectrum tends to be incomplete with the generation of a derivatized residue, thus indicating the presence of a small frequency shift in the mixture. A complete picture of the mutual interactions in xylol is obtained from the subtraction of the two apparently identical spectra A and B (Fig. 5). Both spectra are of xylol, but A is a computer-synthesized mixture (from the spectra of the components) corresponding in composition to a true xylol mixture (spectrum B). The derivative results produced (C) are significant and correlate with frequency shifts of between 0.5 and 1.0 cm^{-1} for each component based on typical band shift data (see Fig. 2B). In this particular example, the experimental conditions were carefully controlled to ensure that no external factors such as cell or instrumental effects could influence the result. Spectrum D (Fig. 5) confirms that these precautions were adequate, since there is no significant difference between two spectra of the same xylol mixture. The lack of any band derivative clearly indicates that the mutual interaction is real.

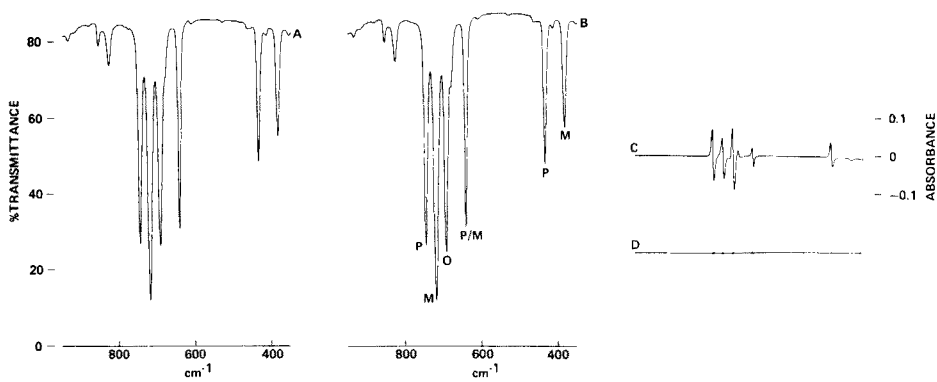


Fig. 5. Component interaction in liquid mixtures illustrated for (*o*, *m*, *p*) xylenes. (A) Computer-synthesized mixture; (B) normal physical mixture; (C) computer difference for B—A; (D) computer difference between two runs of B.

Trace analysis in solution

Frequently samples are submitted for analysis where the analyte is a trace component of a solution. For such samples it is usually necessary to carry out preconcentration procedures, such as solvent extraction, steam or fractional distillation, etc. Under these circumstances a chromatographic technique is often selected as a convenient compromise. If full advantage is taken of the ratio-recording sensitivity of the 580 and the versatility of the SPECT software, i.e. the use of sub-routines for difference, signal averaging, smoothing, etc., it is possible to detect and even identify (under certain circumstances) materials present at trace concentration in solution.

Low levels of hydrocarbon in solution

A classic application of infrared spectroscopy is the determination of trace quantities of hydrocarbon and other organic pollutants present in waste effluent and river water. Initially, the organic material is concentrated by extraction from the aqueous matrix into a small volume of carbon tetrachloride or a similar perhalogenated solvent, usually at a 50:1 water—solvent ratio. The hydrocarbon is determined in solution from the "intense" carbon—hydrogen bands that lie in an area of solvent transparency between 3000 cm^{-1} and 2800 cm^{-1}. Levels of 0.5 ppm or lower in water can be estimated by this procedure when the solutions are measured in 1-cm or 4-cm cells (generally constructed of i.r. grade silica). The overall sensitivity is governed by the degree of pre-concentration in the extraction step and the quality of data produced by the infrared instrument.

Spectrum A of Fig. 6 is a typical result obtained from 20 ppm of hydrocarbon (hexane) in carbon tetrachloride measured in a 4-cm cell, which

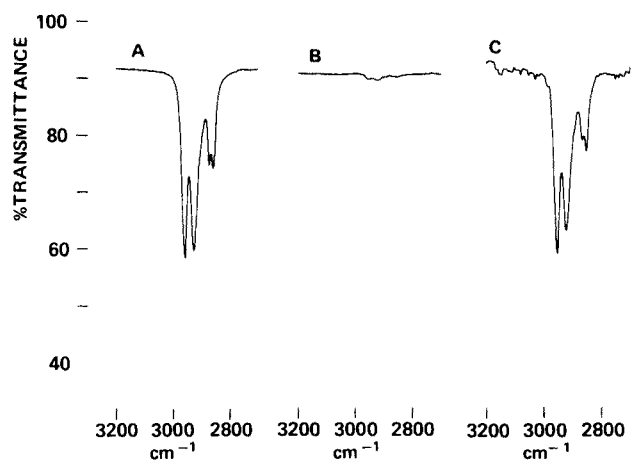


Fig. 6. A: 20 ppm hexane in CCl_4. B: 0.4 ppm hexane in CCl_4. C: Spectrum B expanded to same approx. intensity as spectrum A.

corresponds to 0.4 ppm of the hydrocarbon in water after 50:1 extraction. The same concentration can also be just detected in carbon tetrachloride solution as indicated by spectrum B, although the peak cannot be accurately assessed. With data manipulation on the 580, this peak is easily expanded to the same intensity as the 20-ppm sample. At this level, the concentration is equivalent to 8 ppb (parts per thousand million) in a water sample before extraction. The last spectrum represents a result close to the practical detection limit since further experiments with the 580 indicate that around this level, contamination from glassware and pick-up of hydrocarbon from fingerprints and the environment is critical. This does not, however, represent the ultimate detection of the 580 in this system, which has the potential to detect <2 ppb hydrocarbon in water.

Antioxidants in hydrocarbon media

In the previous example, the procedure lends itself to the detection of trace quantities of material but gives little qualitative data other than the presence of an aliphatic compound. It is often necessary, however, to obtain more information about a component frequently dissolved in a solvent less convenient than carbon tetrachloride. Although the extremely high sensitivity of the previous method cannot be approached, it is possible to utilize the computer software to remove a major solvent matrix (by difference) and expand any residual data by a relatively large factor. As an illustration, the spectrum of a dilute hexane solution of a compound (2, 6-di-t-butyl-*p*-cresol, BHT), used extensively as an antioxidant in oils, foods, polymers and many other consumer and industrial products, is compared with the spectrum of hexane (Fig. 7). The presence of the solute makes only minor modifications to the solvent spectrum as indicated by the arrowed area. Complete removal of the solvent matrix is possible, except in areas of total solvent absorption, i.e. 1485 cm^{-1}—1435 cm^{-1}/1380 cm^{-1}—1360 cm^{-1}, to reveal a low intensity spectrum of the solute. A large data expansion (40× in absorbance) followed by a small degree of smoothing (13 point) produces the characteristic spectrum of the antioxidant (spectrum C). In this particular example, a small degree of signal averaging was used to reduce the influence of noise on the final spectrum. The result clearly demonstrates that good data are obtainable from dilute solutions without extensive preconcentration steps. Typically, it is well suited to the analyses of extracts from aqueous media and eluents in solution from chromatographic separations.

Analysis of blended lubricants

So far, the examples have been selected to illustrate processing features of the 580 CDS system for general requirements in solution studies. The present section deals with a specific area of application that utilizes some of these features, i.e. the lubricating oil industry where most samples can be described as solutions of additives in base oils. There are numerous applications of computerized infrared spectroscopy in the petroleum-based industries and it

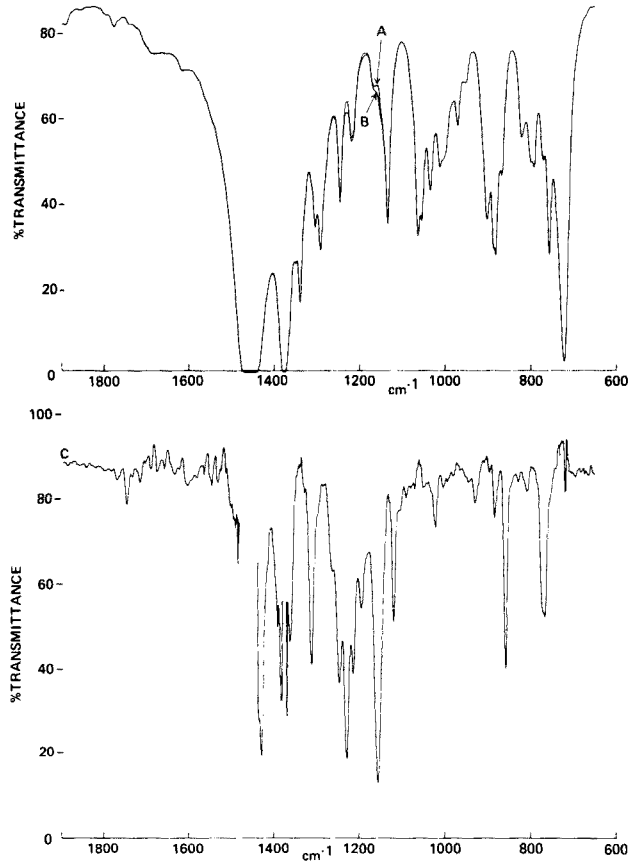


Fig. 7. Analysis for additives in dilute solution. (A) Hexane; (B) 0.1% BHT in hexane; (C) computer difference (B — A) after 40× expansion and digital smoothing (13 pt).

is difficult to do justice to the full potential of the technique. As a compromise, three well-defined areas of interest, that cover typical analytical requirements will be described.

Subtraction of the base oil. One of the first problems that the oil analyst experiences is the difficulty of observing data from the additives blended in a lubricant. Infrared spectroscopy is a good example because most of the spectral data observed originate from the base fluid, usually a hydrocarbon oil. A typical case is recorded in Fig. 8 which illustrates the superimposed spectra of a standard marine lubricant and the base oil used in the lubricant. While some of the additive bands are observed in isolated regions of the spectrum, others are relatively ill-defined and are partially obscured by the base oil absorptions. The difference spectrum, however, produced by subtraction of the base oil after expansion (10× in absorbance) and smoothing (13 point) gives a well-defined spectrum of the additive package. Data in this form can

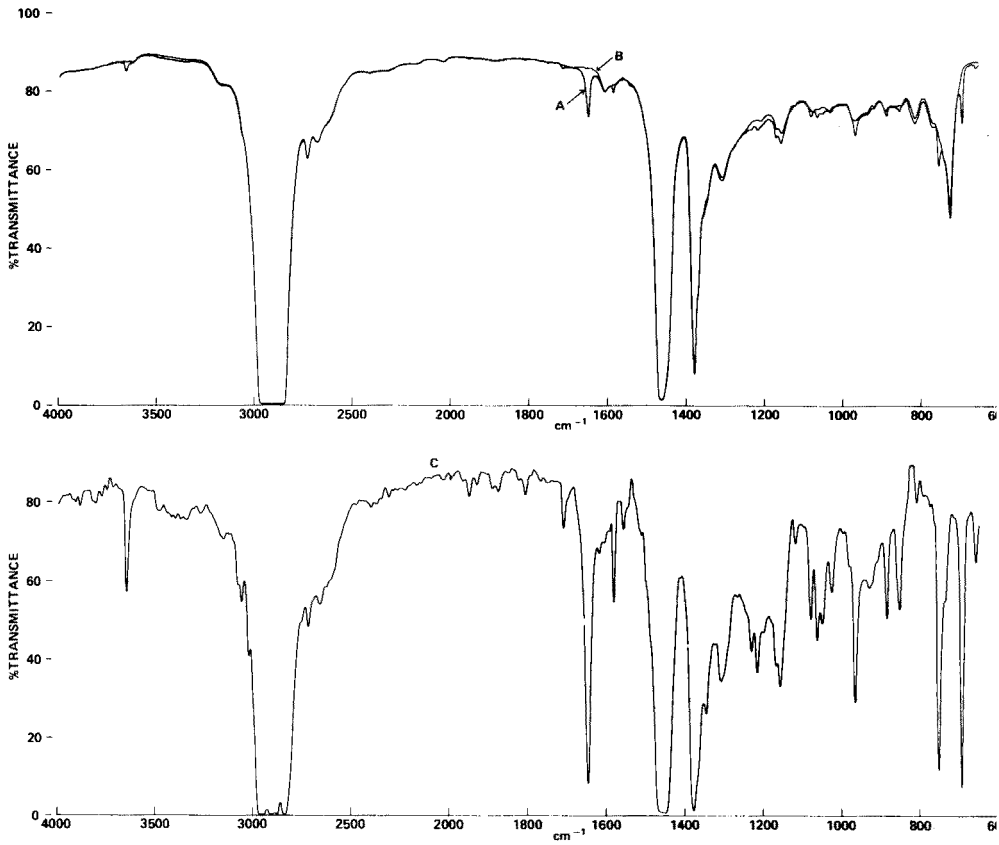


Fig. 8. Base oil extraction from a blended lubricant. (A) Blended lubricant (marine product); (B) base oil; (C) expanded (10) autoscaled difference spectrum equivalent to blended additiv

be used readily to indicate changes in additive formulation and give further data on degradation processes after the oil has been in service.

Detection of blended contaminants. The techniques described above can be extended to the analysis of lubricants at low concentration. The detection of trace components is very important for two reasons: the components may be present as contaminants and as a result impair the performance of the product; a particular additive may only function efficiently at low concentratio A common problem in large-scale blending plants is cross-contamination of blends from additive or lubricant residues retained in pipelines or blending vats. This can cause concern, especially if the product and contaminant are incompatible. Figure 9 shows two superimposed spectra of (A) a suspect, contaminated blend and (B) a standard control blend. The spectral differences are small and could pass without notice if the spectra were not directly compared. In terms of lubricant performance, however, these differences are important. After subtraction of the major background, i.e. spectrum B, and a

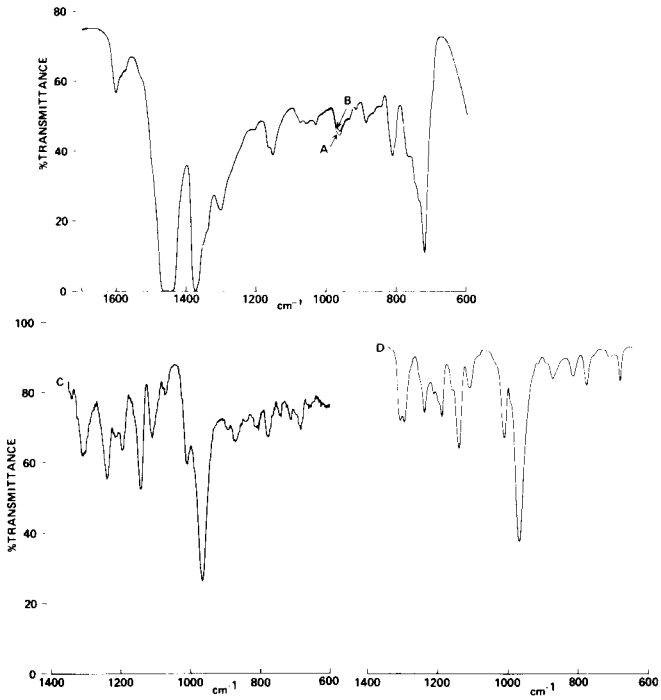


Fig. 9. Detection of trace (or contaminant) additives in lubricating oils. (A) Standard lubricant (200-μm pathlength cell); (B) sample A with 400 ppm phosphate additive; (C) computer difference (A — B) with scale expansion and digital smoothing; (D) reference spectrum of phosphate additive (tricresyl phosphate).

large data expansion, approximately 40X in concentration, with smoothing, a spectrum of the contaminant is obtained (spectrum C). Comparison of these spectra with standard spectra reveals that the material is an aryl phosphate, probably tricresyl phosphate (spectrum D) and is estimated to be present at ca. 400 ppm. The experiment was repeated with signal averaging for different samples, and results indicated that under the same basic conditions it is possible to detect 100 ppm of the contaminant in the same lubricant.

Lubricant quality control. A further problem in lubricant production is the quality control of oil blends based on multicomponent formulations containing complex additive packages. In Fig. 10, A and B are the spectra of two similar additive packages incorporating several additives for engine cleanliness and wear/corrosion protection. Both packages are produced for marine diesel lubricants. There are obvious, yet subtle, differences in the two spectra, which are impossible to observe in the fully formulated blend. This fact is clearly confirmed in Fig. 11, where spectrum A is produced from the correct formulation containing both additives in equal proportion and spectrum B is a similar blend modified with a slight excess of one of the packages. The standard

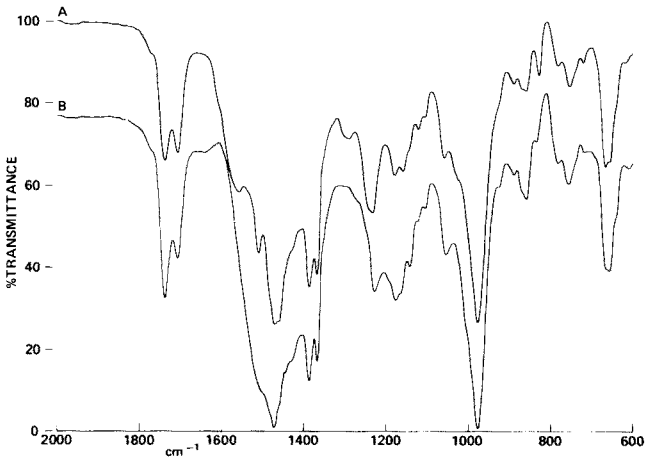


Fig. 10. Oil additive packages. Data produced from lube oil blend (10% additive) after computer subtraction of base oil. (A) Additive package X; (B) Additive package Y.

formulation can be used as a control, and subtraction of its spectrum (A) from the spectrum of any suspect product, such as B, results in a difference spectrum for correlation with the standard spectra of the blended constituents. In the case of spectrum B, the residual difference after 10× concentration expansion and digital smoothing (Fig. 11C) compares well with spectrum A of Fig. 10 to pinpoint additive X present in excess relative to the normal product specification.

The success of the above exercise is very significant, for it provides a suitable method of matching standard and production lubricants. Any degree of mismatch indicates a deviation in formulation and the data produced may be utilized for quality control of the additives and the final product.

Study of samples in aqueous media

A major part of routine infrared analysis in solution is limited to studies in non-aqueous, usually organic solvents. Many samples submitted for analysis, however, are not readily soluble in organic media and have to be examined as a solution in water. The study of this type of sample is normally restricted by the following factors: water absorbs infrared radiation throughout most of the analytical region; common infrared window materials are hygroscopic; the sensitivity of infrared instrumentation is frequently inadequate for the analysis of dilute solutions. The hygroscopicity of normal window materials can be a genuine limitation, because water-insoluble windows either have a very restricted infrared transmission range or are very expensive. Barium fluoride gives a reasonable compromise between cost and transmission for data down to 800 cm^{-1}. The other two factors are easily overcome with the aid of the 580 and computer data processing. Several new applications, based on samples prepared as solutions in water, have been completed with the CDS-

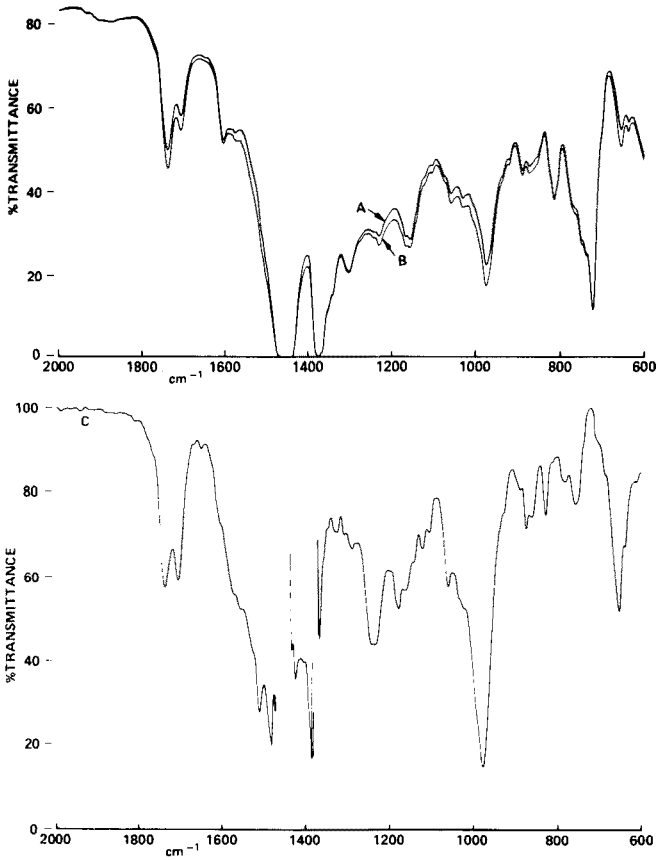


Fig. 11. Quality control of lubricating oils. (A) Laboratory standard blend; (B) production blend with slight excess of one package additive; (C) expanded difference spectrum (B — A) — compare with reference data (Fig. 10).

(580) system. Two examples, chosen to illustrate the flexibility of the system for work in an aqueous environment, are discussed here.

Many commercial products are manufactured and processed as emulsions or suspensions in water, e.g. synthetic rubber latex. The infrared analysis of this type of material is usually performed on the free polymer produced by evaporation of the water. Often the results obtained do not reflect the composition of the aqueous product because volatile components are often lost during the evaporation stage, e.g. monomers, stabilizers, auxiliary solvents, etc., and because many polymers are prone to oxidation or even cross-linking in the isolated state. It is, therefore, advantageous to examine the complete product and remove the interference of water by data processing. Figure 12(A) is the spectrum of a styrene—butadiene copolymer latex in a short pathlength, barium fluoride cell. The most intense features in this spectrum are attributed

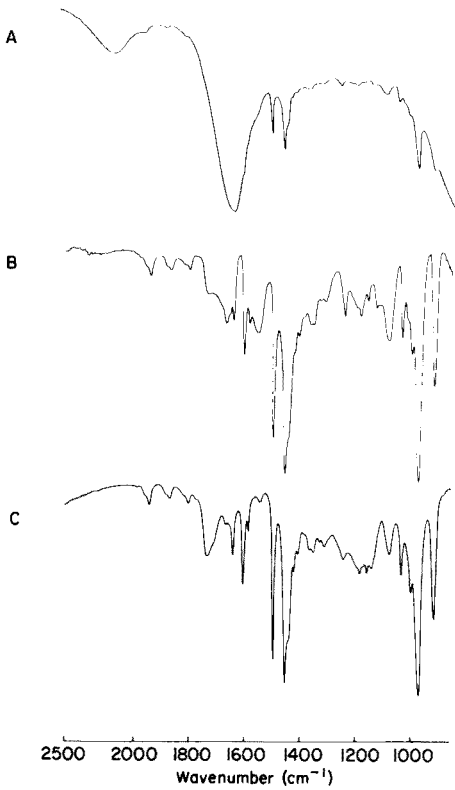


Fig. 12. Aqueous synthetic latex emulsion. (A) Solution spectrum (7-μm BaF_2 cell); (B) water-subtracted spectrum, 4X ordinate expansion; (C) isolated polymer spectrum evaporate on silver bromide.

to water. After computer subtraction of the water background and 4X data expansion (in absorbance), a clearly defined spectrum of the emulsified material is produced (spectrum B). For interest, it is worth comparing the processed data with the spectrum acquired from an evaporated film of the latex. Minor spectral differences are observed in the 1730 cm^{-1} and 1100 cm^{-1} regions of the evaporated film spectrum. These variations are attributed to partial oxidation of some unsaturated sites on the isolated polymer.

The second application relates to a problem encountered in the food industr Food and drink products often contain a mixture of natural and synthetic constituents and under modern consumer laws it is important to be able to differentiate these two types of ingredient. One example is the use of non-dairy products, especially in beverages, as a substitute for milk or cream. Infrared spectroscopy would not usually be considered for this application because the data produced do not adequately distinguish the two products. The problem is well illustrated by the spectra in Fig. 13 of a coffee drink

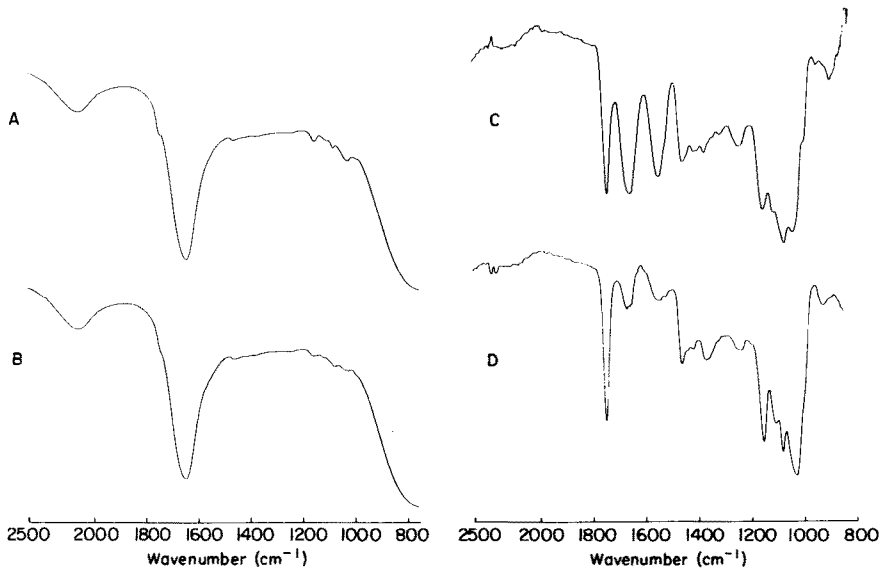


Fig. 13. Comparison of natural and synthetic food products in aqueous solution. (A) Coffee and milk; (B) coffee and non-dairy creamer; (C) coffee and milk, expanded approx. 20×; (D) coffee and non-dairy creamer, expanded approx. 20×.

containing (A) milk and (B) a milk substitute. Some indication of dissolved material is given but the water absorptions mask most of the spectral details, making it impossible to differentiate between the two beverages. The situation changes when the interference from water is removed by computer difference. Although relatively weak, the residual data may be expanded to give spectra C and D. Spectral differences between the milk and the milk substitute are now obvious and are equated to the presence of fats (animal in the milk/vegetable in the substitute), protein (in the milk), sugars (in both milk and substitute) and inorganic phosphate and/or amorphous silica in the substitute (additional bands around 1000 cm^{-1}).

The diversion of infrared spectroscopy into routine analysis in aqueous media with the aid of data processing leads to many new applications. A more detailed account of this technique has been given [2].

Conclusions

The application of computers to the processing of spectroscopic data is gaining popularity now that computer-compatible instruments and low-cost data systems are available. Some illustrations of how the technique may be applied at a fairly basic level to generate data that are not usually readily accessible from conventional spectrophotometers have been discussed here.

Over the next few years, it is expected that this type of computerization will expand to such an extent as to form the basis of many routine analytical experiments.

REFERENCES

1 D. J. Gardiner, R. B. Girling and R. E. Hester, J. Chem. Soc., Faraday Trans. 2, 71 (1975) 709
2 J. P. Coates, European Spectroscopy News, (1978) 16, 25.

Analytica Chimica Acta, 103 (1978) 339—354
Computer Techniques and Optimization

THEORY OF ERROR FOR TARGET FACTOR ANALYSIS WITH APPLICATIONS TO MASS SPECTROMETRY AND NUCLEAR MAGNETIC RESONANCE SPECTROMETRY

EDMUND R. MALINOWSKI

Department of Chemistry and Chemical Engineering, Stevens Institute of Technology, Hoboken, New Jersey 07030 (U.S.A.)

(Received 3rd May 1978)

SUMMARY

Based on the theory of error for abstract factor analysis described earlier, a theory of error for target factor analysis is developed. The theory shows how the error in the data matrix mixes with the error in the target test vector. The apparent error in a target test is found to be a vector sum of the real error in the target vector and the real error in the predicted vector. The theory predicts the magnitudes of these errors without requiring any a priori knowledge of the error in the data matrix or the target vector. A reliability function and a spoil function are developed for the purpose of assessing the validity and the worthiness of a target vector. Examples from model data, mass spectrometry and nuclear magnetic resonance spectrometry are presented.

During the present evolutionary stages of chemometrics [1], factor analysis has been shown to be a powerful tool for solving multidimensional problems in chemistry [2, 3]. Factor analysis attempts to express the points in a data matrix as a linear sum of product terms called factors [4—8]. The first stages of factor analysis involve determining the number of controlling factors and reproducing the data by means of these mathematical abstract factors. Unfortunately, experimental error tends to confuse the process at this early stage. Recently, Malinowski [9] developed a theory of error for abstract factor analysis which shows how the error mixes into the general scheme. This theory makes it possible to determine the dimensionality of the factor space without any a priori knowledge of the experimental uncertainty. In fact, the theory can be used to estimate the real error in the data. Furthermore, it shows that abstract factor compression actually leads to data improvement.

The second stage of factor analysis concerns the transformation of the abstract eigenvector axes, which emerge from the first stage, into real axes which have chemical significance. By means of target rotation [10, 11] each real axis can be sought individually and independently of all other axes.

Unfortunately, target testing is complicated not only by the errors in the data matrix but also by the errors in the target itself. Currently, the decision as to whether or not a given target is a valid representation of a real factor rests too heavily on the intuitive judgement of the analyst carrying out the target test. The commonest method is to compare, point by point, the elements of the target with the corresponding elements of the predicted vector. If the vectors look alike, the target is accepted; if not, the target is rejected. Because this approach relies entirely on the subjective whim of the analyst, incorrect conclusions can too readily be reached. More exacting methods, which are free from such bias, are sorely needed for target factor analysis.

This paper reports an attempt to develop quantitative criteria for target testing. How the errors from the data matrix and the target test mix together, and how a knowledge of this mixing phenomenon yields valuable information concerning the true controlling factors is demonstrated. The approach is based on the theory of error for abstract factor analysis, which is extended to target factor analysis. The general development, arguments, notation and derivations depend heavily on the earlier investigation [9], which is therefore reviewed briefly below.

Synopsis of target factor analysis

The first step of factor analysis involves decomposing the data matrix [D] into two abstract matrices [R] and [C]: [D] = [R][C]. This means that each data point, d_{ik}, is expressed as a linear sum of factors, each factor being a product of a row cofactor, r_{ij}, and a column cofactor, c_{jk}:

$$d_{ik} = \sum_{j=1}^{j=n} r_{ij}\, c_{jk} \qquad (1)$$

The subscripts specify the locations (row and column) of the elements in their respective matrices. The sum is taken over n principal factors which account for the data within experimental error. The size of [D] is $r \times c$ where r is the number of rows and c is the number of columns in the data matrix. [R] is an $r \times n$ matrix and [C] is an $n \times c$ matrix. [R] is called the row-factor matrix because its elements are associated with the row designees of the data matrix. [C] is called the column-factor matrix because its elements are associated with the column designees of the data matrix. Numerical values for the elements of [R] and [C] are routinely obtained from [D] by applying the mathematical procedure known as principal component analysis; this can be done efficiently by computer. Several specially designed computer programs are currently available [12, 13].

Unfortunately, [R] and [C] represent mathematical solutions which are devoid of physical or chemical meaning. In factor analysis these abstract matrices are transformed to significant real matrices by a transformation matrix [T]. When [R] is postmultiplied by the appropriate [T], a physically significant row-factor matrix, $[\bar{R}]$, emerges: $[\bar{R}] = [R][T]$. It is mathematically apparent that if [C] is premultiplied by the inverse of the same transformation

matrix, the associated column matrix, $[\bar{C}]$, emerges: $[\bar{C}] = [T]^{-1}[C]$. Thus, the data matrix can be expressed in terms of real quantities: $[D] = [\bar{R}][\bar{C}]$.

The procedure for obtaining the appropriate [T] is called target testing [10, 11]. Each column of [T] can be obtained independently of the other columns. If $\bar{R}_l$ represents the l-th column of $[\bar{R}]$ and T_l represents the analogous column of [T] then,

$$\bar{R}_l = [R]T_l \tag{2}$$

In other words, each column of [T] is associated with only one of the real factors responsible for the data.

Symbolically, $\bar{\bar{R}}_l$ is defined as representing the l-th test vector composed of numerical values for each of the r designees. For example, if the row designees were the normal hydrocarbons ranging from methane to octane and the boiling point was suspected to be a factor, then $\bar{\bar{R}}_l$ would be a test vector composed of the eight respective boiling points. Of course, it is necessary to seek a transformation vector, T_l which yields a "predicted vector", $\bar{R}_l$, closely matching $\bar{\bar{R}}_l$. Using a method of least squares, Malinowski [10] showed that such a transformation vector can be calculated from the abstract row matrix:

$$T_l = \{[R]^T[R]\}^{-1}[R]^T\bar{\bar{R}}_l \tag{3}$$

This expression can be appropriately modified to accommodate the situation where the test vector is incomplete.

Eigenvector interpretation

Each column of [D] can be looked on as an axis associated with the column designee. These data axes are not necessarily orthogonal. Instead of lying in c space they lie in n space, i.e. only n axes (eigenvectors) are required to describe the situation; these eigenvectors are obtained by the principal component analysis procedure. Because of experimental error, c eigenvectors are always obtained if $c < r$. However, only n eigenvectors, associated with the largest eigenvectors, called the primary set, are required to describe the factor space. The remaining set of eigenvectors — the secondary set — is composed solely of experimental error, contains no useful information and is discarded from further consideration. The primary set of eigenvectors is used to construct [C]. Each column of [C] is a primary eigenvector.

Each row designee of [D] can be looked on as a point in the factor space. The perpendicular projection of a row-designee point onto a column-designee axis gives the value of the designated data point in [D]. The perpendicular projection of a row-designee point onto an eigenvector axis gives the value of the designated point in [R]. [C] is composed of eigenvectors and [R] is composed of the projections of the row-designees onto the respective eigenvectors. An eigenvalue is the sum of the squares of the projections of all of the row-designee points onto the respective eigenvector axis. The j-th eigenvalue associated with the j-th eigenvector is simply

$$\lambda_j = \sum_{i=1}^{i=r} r_{ij}^2 \tag{4}$$

Effect of experimental error

Because of experimental uncertainty, a data point is best represented as a sum of a pure value, d_{ik}^*, and an error, e_{ik}:

$$d_{ik} = d_{ik}^* + e_{ik} \tag{5}$$

According to the theory of error for abstract factor analysis [9], deletion of the secondary set of eigenvectors reduces the error in the data reproduced by abstract factor analysis, $d_{ik}^{\neq}$, so that $d_{ik}^{\neq} = d_{ik}^* + e_{ik}^{\neq}$, where $e_{ik} = e_{ik}^{\neq} + e_{ik}^0$. Here e_{ik}^0 is the amount of error which is effectively removed by deleting the secondary eigenvectors and $e_{ik}^{\neq}$ is the amount of error which remains in the reproduced data after the deletion. These equations can be written in matrix form: [D] = [D*] + [E]; [D$^{\neq}$] = [D*] + [E^0]; and [E] = [E$^{\neq}$] + [E^0]. If the data were pure, a solution could be obtained in terms of the pure matrices: [D*] = [R*][C*]

In chemistry, such situations are rarely, if ever, encountered. Instead the data matrix will normally contain some experimental error and, consequently, the resulting row-factor and column-factor matrices will contain some error. Hence, [D$^{\neq}$], the data matrix reproduced by abstract factor analysis, will differ from [D].

$$[D^{\neq}] = [R][C] \tag{6}$$

Malinowski [9] postulated that the same basic axes used to describe the raw data space can be used to describe the error space. However, instead of using only the primary set, the entire set of eigenvectors must be used to describe the complete error space. Accordingly, the error may be expressed as

$$e_{ik} = \sum_{j=1}^{j=n} \sigma_{ij}^{\neq} c_{jk} + \sum_{j=n+1}^{j=c} \sigma_{ij}^0 c_{jk} \tag{7}$$

Here $\sigma_{ij}^{\neq}$ is the projection of the i-th error onto the j-th primary axis and σ_{ij}^0 is the projection of the i-th error onto the j-th secondary axis. Placing eqn. (7) into eqn. (5) and applying the form of eqn. (1) to the pure data gives

$$d_{ik} = \sum_{j=1}^{j=n} (r_{ij}^* c_{jk}^* + \sigma_{ij}^{\neq} c_{jk}) + \sum_{j=n+1}^{j=c} \sigma_{ij}^0 c_{jk} = \sum_{j=1}^{j=n} r_{ij}^{\neq} c_{jk} + \sum_{j=n+1}^{j=c} \sigma_{ij}^0 c_{jk} \tag{8}$$

where

$$r_{ij} = r_{ij}^* (c_{jk}^*/c_{jk}) + \sigma_{ij}^{\neq} \tag{9}$$

Here, r_{ij}^* is an element of [R*], r_{ij} is an element of [R], c_{jk}^* is an element of [C*], and c_{jk} is an element of [C].

Examination of eqns. (4), (8) and (9) shows that

$$\lambda_j^{\neq} = \sum_{i=1}^{i=r} r_{ij}^2 \text{ for } j = 1, \ldots, n \tag{10}$$

and

$$\lambda_j^{0} = \sum_{i=1}^{i=r} (\sigma_{ij}^{0})^2 \text{ for } j = n+1, \ldots, c \tag{11}$$

where $\lambda_j^{\neq}$ represents an eigenvalue associated with a primary eigenvector and λ_j^{0} represents an eigenvalue associated with a secondary eigenvector. These are the eigenvalues which emerge from factor analysis of the raw data matrix which contains experimental uncertainty.

The residual standard deviation (RSD) is defined by the standard equation:

$$r(c-n)\,(\text{RSD})^2 = \sum_{i=1}^{i=r} \sum_{j=n+1}^{j=C} (\sigma_{ij}^{0})^2 \tag{12}$$

Insertion of eqn. (11) into eqn. (12) and rearrangement gives

$$\text{RSD} = \left[\frac{\sum_{j=n+1}^{j=C} \lambda_j^{0}}{r(c-n)}\right]^{1/2} = \text{RE} \tag{13}$$

From the secondary eigenvalues the RSD, also known as the real error RE, can be easily calculated. The RE represents the root-mean-square of the errors projected onto all of the secondary axes. If the errors are randomly scattered throughout the data matrix, the same result should be obtained by calculating the root-mean-square of the errors projected onto all of the primary axes. This means that the RSD in eqn. (13) should be identical to the RSD expressed by

$$rn(\text{RSD})^2 = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (\sigma_{ij}^{\neq})^2 \tag{14}$$

This hypothesis is used in the development of a theory of errors for target transformation.

THEORY OF ERROR FOR TARGET TRANSFORMATION

Because of experimental error, target testing is not a simple, straightforward process. Complications arise because errors from two sources, the data matrix and the target vector itself, combine in the testing procedure. The target testing process is investigated below in order to learn how these errors combine. The approach is based on the theory of error summarized in the preceding paragraphs. Separation of these errors and estimation of their magnitudes can be achieved without any a priori knowledge of the errors. Two criteria, the reliability function and the spoil function, are developed for assessing the validity and the worthiness of a target vector.

Theoretical derivation

Equation (9) shows how the errors in the data matrix perturb the abstract row cofactors. If these errors are reasonably small, c_{jk}^*/c_{jk} will be close to unity, and eqn. (9) reduces to $r_{ij} = r_{ij}^* + \sigma_{ij}^{\neq}$. Thus

$$[R] = [R^*] + [E^{\neq}] \tag{15}$$

Here $[E^{\neq}]$ is an error matrix composed of $\sigma_{ij}^{\neq}$; $[R^*]$ is the pure row-factor matrix which can be discussed theoretically but can never really be obtained. $[R]$ is the abstract row-cofactor matrix which automatically emerges during the first stage of factor analysis; thus the numerical values of all of the elements of this matrix are accessible. It is this matrix $[R]$ that must be used in conjunction with eqn. (3) to obtain the transformation vector; and this matrix must also be used, with eqn. (2), to obtain the predicted vector.

The difference between the predicted test vector and the target vector is defined as the "apparent" error vector $\mathbf{E}_A$,

$$\mathbf{E}_A = \bar{\mathbf{R}}_l - \bar{\bar{\mathbf{R}}}_l \tag{16}$$

Combination of eqns. (2), (15) and (16) gives

$$\mathbf{E}_A = [E^{\neq}]\ \mathbf{T}_l + [R^*]\ \mathbf{T}_l - \bar{\bar{\mathbf{R}}}_l \tag{17}$$

If the data matrix and the test vector contained no experimental error and if the test vector were a true factor, then the predicted vector, $\bar{\mathbf{R}}_l^*$, would be pure and would exactly equal the pure test vector, $\bar{\bar{\mathbf{R}}}_l^*$; thus $\bar{\bar{\mathbf{R}}}_l^* = \bar{\mathbf{R}}_l^* = [R^*]\ \mathbf{T}_l^*$ where $\mathbf{T}_l^*$ is the transformation vector for the l-th test vector, obtained from pure data. In chemistry, pure data are rarely, if ever, attainable; consequently, impure data must be used. If the errors are small, the following approximation should hold: $\bar{\bar{\mathbf{R}}}_l^* \cong [R^*]\ \mathbf{T}_l$ where $\mathbf{T}_l$ is the transformation vector generated from impure data. If this were true, then eqn. (17) would become

$$\mathbf{E}_A \cong [E^{\neq}]\ \mathbf{T}_l + \bar{\bar{\mathbf{R}}}_l^* - \bar{\bar{\mathbf{R}}}_l \tag{18}$$

The difference between the pure target vector, $\bar{\bar{\mathbf{R}}}_l^*$, and the raw target vector is called the "real error in the target test vector", $\mathbf{E}_T$,

$$\mathbf{E}_T = \bar{\bar{\mathbf{R}}}_l^* - \bar{\bar{\mathbf{R}}}_l \tag{19}$$

The difference between the predicted vector and the pure target vector is called the "real error in the predicted vector", $\mathbf{E}_p$,

$$\mathbf{E}_p = \bar{\mathbf{R}}_l - \bar{\bar{\mathbf{R}}}_l^* \tag{20}$$

From the definitions given by eqns. (16), (19) and (20):

$$\mathbf{E}_A = \mathbf{E}_p + \mathbf{E}_T \tag{21}$$

Inserting eqn. (19) into eqn. (18) and comparing the result with eqn. (21) leads to

$$\mathbf{E}_p = [E^{\neq}]\ \mathbf{T}_l \tag{22}$$

It is common practice to deal with root-mean-square (RMS) errors rather

than error vectors. The RMS of the apparent error in the target test vector (AET) is related to the inner product of the error vector and is defined by

$$\mathbf{E}_A^T \mathbf{E}_A = \sum_{i=1}^{i=r} (\bar{r}_i - \bar{\bar{r}}_i)^2 = r(\mathrm{AET})^2 \tag{23}$$

In this expression, the inner product is equal to the error vector, $\mathbf{E}_A$, premultiplied by its own transpose, $\mathbf{E}_A^T$; $\bar{r}_i$ and $\bar{\bar{r}}_i$ are the i-th elements of the predicted vector and target vector, respectively; the sum is taken over all r elements of the vector. Similarly, the root-mean-square of the real error in the predicted vector (REP) is defined by expression:

$$\mathbf{E}_P^T \mathbf{E}_P = \sum_{i=1}^{i=r} (\bar{r}_i - \bar{\bar{r}}_i^*)^2 = r(\mathrm{REP})^2 \tag{24}$$

where $\bar{\bar{r}}_i^*$ is the i-th element of the pure target vector. Finally, the root-mean-square of the real error in the target vector (RET) is defined by

$$\mathbf{E}_T^T \mathbf{E}_T = \sum_{i=1}^{i=r} (\bar{\bar{r}}_i^* - \bar{\bar{r}}_i)^2 = r(\mathrm{RET})^2 \tag{25}$$

To see the interrelationship between these three root-mean-square errors, the inner product of both sides of eqn. (21) is first required:

$$\mathbf{E}_A^T \mathbf{E}_A = \mathbf{E}_P^T \mathbf{E}_p + \mathbf{E}_T^T \mathbf{E}_T + \mathbf{E}_P^T \mathbf{E}_T + \mathbf{E}_T^T \mathbf{E}_P \tag{26}$$

Since the elements of these vectors are random errors scattered about zero, being both positive and negative, the final two terms on the right should be small in comparison to the first two terms on the right, which are sums of squares. Thus

$$\mathbf{E}_A^T \mathbf{E}_A = \mathbf{E}_P^T \mathbf{E}_P + \mathbf{E}_T^T \mathbf{E}_T \tag{27}$$

Substitution of eqns. (23—25) into eqn. (27) gives, by definition,

$$(\mathrm{AET})^2 = (\mathrm{REP})^2 + (\mathrm{RET})^2 \tag{28}$$

Figure 1 is a mnemonic diagram depicting this pythagorean relationship. For a given target test, the three vectors, $\bar{\bar{\mathbf{R}}}$, $\bar{\bar{\mathbf{R}}}^*$ and $\bar{\mathbf{R}}$, lie at the corners of the right triangle: the pure vector lies at the right angle, the predicted vector at the acute base angle, and the impure target vector at the apex. The hypotenuse represents AET; the base represents REP; and the height represents RET.

Evaluation of AET, REP and RET

In many instances, a complete description of the target cannot be made because of lack of information or data. Instead of all r elements, only p elements of the target may be available. In spite of this shortcoming, the apparent error can be estimated by considering the number of degrees of freedom. From statistical considerations:

$$\left[\sum_{i=1}^{i=r} (\bar{r}_i - \bar{\bar{r}}_i)^2\right]/[r-n] \cong \left[\sum_{i=1}^{i=p} (\bar{r}_i - \bar{\bar{r}}_i)^2\right]/[p-n] \tag{29}$$

Combining eqns. (23) and (29) gives a general expression for estimating the apparent error:

$$\mathrm{AET} = \left[\frac{r-n}{p-n}\,\frac{\Sigma_{i=1}^{i=p}(\bar{r}_i - \bar{\bar{r}}_i)^2}{r}\right]^{1/2} \tag{30}$$

For numerical evaluation of REP, the reasoning is as follows. According to eqn. (22) the inner product of the predicted vector can be written as:

$$\mathbf{E}_P^T\,\mathbf{E}_P = \{[\mathbf{E}^{\neq}]\,\mathbf{T}_l\}^T\{[\mathbf{E}^{\neq}]\,\mathbf{T}_l\} = \mathbf{T}_l^T[\mathbf{E}^{\neq}]^T\,[\mathbf{E}^{\neq}]\,\mathbf{T}_l \tag{31}$$

The error product $[\mathbf{E}^{\neq}]^T\,[\mathbf{E}^{\neq}]$ is an $n \times n$ matrix with a trace equal to $\Sigma_{i=1}^{i=r}\Sigma_{j=1}^{j=r}(\sigma_{ij}^{\neq})^2$. According to eqn. (14), this sum is equal to $rn(\mathrm{RSD})^2$. Hence the trace of the error product matrix is $rn(\mathrm{RE})^2$, where RE is the real error in the data matrix. RE can be calculated from the secondary eigenvalues by means of eqn. (13). The off-diagonal elements of the error product matrix are expected to be negligibly small in comparison to the diagonal elements because $[\mathbf{E}^{\neq}]$ is an error matrix composed of a random set of positive and negative errors scattered about zero. The product matrix has n elements along the diagonal, hence the average value of a diagonal element is $r(\mathrm{RE})^2$. On the basis of this argument, eqn. (31) can be approximated by

$$\mathbf{E}_P^T\,\mathbf{E}_P = r(\mathrm{RE})^2\,(\mathbf{T}_l^T\,\mathbf{T}_l) \tag{32}$$

where $\mathbf{T}_l^T\,\mathbf{T}_l$ is the inner product of the transformation vector. Thus, from eqn. (24),

$$\mathrm{REP} = (\mathrm{RE})\,(\mathbf{T}_l^T\,\mathbf{T}_l)^{1/2} \tag{33}$$

Because numerical values for RE and $\mathbf{T}_l$ are readily available from factor analysis, this equation provides a method for evaluating the real error in the predicted vector.

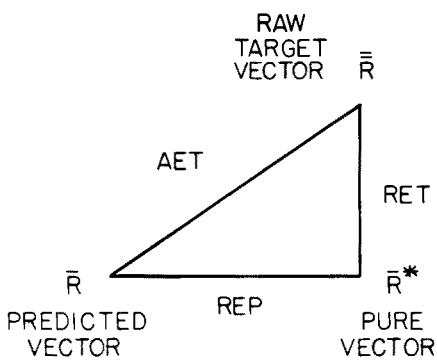


Fig. 1. Mnemonic diagram of the Pythagorean relationship between the apparent error in the target (AET), the real error in the target (RET) and the real error in the predicted target (REP), and their relationship to the pure, raw and predicted target vectors.

After numerical values have been obtained for AET and REP by means of eqns. (30) and (33), respectively, RET can be calculated from eqn. (28), rearranged as

$$RET = [(AET)^2 - (REP)^2]^{1/2} \tag{34}$$

If the calculated RET agrees reasonably well with knowledge of the error in the data used to construct the test vector, it would be concluded that the target is a real factor. If the calculated RET is much greater than the estimated error, the target would be rejected since it would not be a valid representation of a true factor.

The present derivations are based on two major premises: (1) that the target vector is true vector; and (2) that the errors in the data matrix are sufficiently small. The theoretical expressions for AET, REP and RET will not hold if these two conditions are not satisfied. In the following section additional error functions are developed, to help decide whether or not these conditions are fulfilled and, therefore, whether or not a given target is acceptable.

Reliability function

A direct comparison between the calculated and estimated values of RET could be misleading because this involves absolute differences. It is better to consider relative differences. Hence the use of a reliability function, RELI, is proposed; this measures the error difference relative to the apparent error,

$$RELI = \left[1 - \frac{(RET)^2 - (RET)^2_{est}}{(AET)^2}\right]^{1/2} \tag{35}$$

In this expression, $(RET)_{est}$ is the estimated value for the real error in the target.

For a given target, if REP is found to equal AET, then RET is zero and the reliability would be unity, i.e. the upper limit of RELI, indicative of a pure test vector without error. In actual practice, RELI values may be greater than unity, because of the approximations involved in deriving the expressions used to evaluate RET and REP, and because of the possibility of overestimating $(RET)_{est}$ from experimental information. However, if the calculated RET is excessively larger than the estimated RET, then the RELI would be close to zero and no confidence should be placed in the target. As a rule of thumb, if the RELI is above 0.5, the target can be regarded as reliable; if the RELI is below 0.5, the target is probably unreliable. In this sense, the RELI function acts as a gauge which indicates how close the target matches a true factor. It should be stressed that the RELI function is a new concept in factor analysis and its exact meaning and interpretation will be subject to appropriate modifications as dictated by further investigation.

Spoil function

Although a target test vector may have a RELI value close to unity, its utilization in data matrix reproduction may lead to increased error. This will

occur if the test target is a true factor but is composed of data having relatively large error in comparison to the error in the data matrix. The use of a "spoil" function, SPOIL, is proposed here as a criterion to help determine when such a situation arises.

Before the SPOIL function is defined, it is important to examine and reinterpret the REP function. According to eqn. (28), cf. Fig. 1, if a given target vector had no error, then RET would be zero and AET would equal REP. This means that all the error in the predicted vector must come from the data matrix, because error cannot be contributed by a pure target. To emphasize this important point, we can write REP = EDM, where EDM is the error in the predicted vector contributed by the "error from the data matrix". These arguments lead to the conclusion that the apparent error in the target is a vector sum of two errors which originate from two distinct sources: (1) an error from the impure data matrix and (2) an error from the impure target data.

This interpretation provides an excellent insight into the target process. For example, if the real error in a target is smaller than the error from the data matrix, then the error in the reproduced target will be larger than the error in the original target, i.e. the reproduced target will be less reliable than the original target. This becomes evident if Fig. 1 is examined carefully. If the target data are more accurate than the data matrix, their usage in the target combination step will lead to improvement in the regenerated data matrix. Hence, to describe the target the most accurate data possible should be used.

Conversely, if the real error in a target is greater than the error from the data matrix, then the error in the reproduced target will be less than the error in the original target, i.e. the reproduced target will be more accurate than the original target. This offers an unexpected fringe benefit; target testing can be used as a particular means of improving data, provided that the target is a true factor and that the correct number of eigenvectors is employed in the transformation process. One important point should be made here. The method does not require identification of the other factors involved. Unfortunately, however, when such targets are employed in the data matrix reproduction step, the additional errors from the targets mix into the process and spoil the reproduced data.

The SPOIL function is designed to measure the increase or decrease in the error associated with the reproduced data matrix which would be caused by replacing an abstract eigenvector with a given target vector. SPOIL is defined as the ratio of the real error in the target and the error from the data matrix.

$$\text{SPOIL} = \text{RET}/\text{EDM} = \text{RET}/\text{REP} \tag{36}$$

The reproduced data matrix will be improved by a target with a SPOIL less than 1.00 and ruined by a target with a SPOIL greater than 1.00. In fact, the larger the SPOIL, the worse will be the reproduction. Hence, accurate targets with SPOIL values close to zero should be sought.

SPOIL has another valuable aspect. There is often no way of estimating the amount of error in the target and, consequently, its reliability cannot be

estimated from the RELI function. In such situations, the SPOIL function can be used as a crude measure of the acceptability of the target. Studies of model sets of data have indicated a rule-of-thumb criterion which relates the SPOIL to the degree of acceptability. The SPOIL values shown in Fig. 2 are somewhat arbitrarily divided into three general regions: (1) an acceptable region (0.0—3.0); (2) a moderately acceptable region (3.0—6.0); and (3) an unacceptable region (>6.0). The first two regions can be further subdivided into SPOIL regions which provide a physical and verbal sense of the overall validity of the target. If the SPOIL is 0.0, the target is considered perfect; this is the ultimate level of achievement seldom experienced. The target is considered excellent for SPOIL ≤ 1.5; a SPOIL between 1.5 and 3.0 is considered good, and so on, as indicated in Fig. 2.

Obviously, the SPOIL function is a new concept, and its interpretation and usage will be subject to modification as dictated by future investigations.

MODEL STUDIES

Before the new error theory is applied to real chemical data, it is important to test the theory by using model sets of data for which all information is known. Artificial data matrices were created by first assigning numerical values to the elements of $[\bar{R}^*]$ and $[\bar{C}^*]$, the pure row-factor and column-factor matrices. Matrix multiplication then yielded a pure data matrix, to which an artificial, known error matrix was added. The resulting impure data matrix

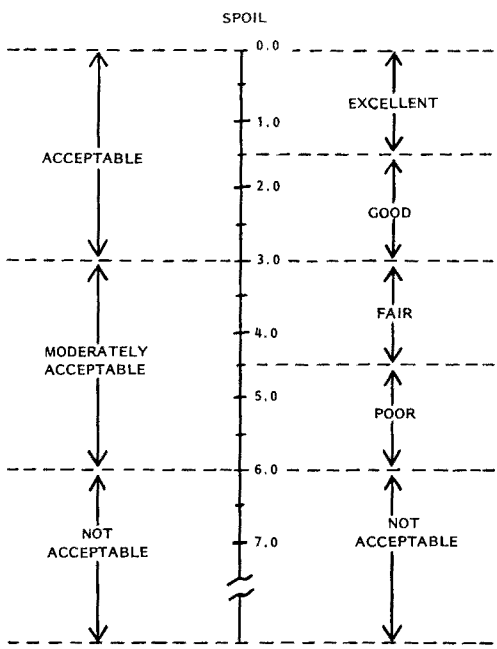


Fig. 2. Summary diagram showing the rule-of-thumb criterion relating the SPOIL to the degree of acceptability.

was meant to simulate real chemical data. A series of impure target vectors were generated by adding different amounts of known error to each column of $[\bar{R}^*]$. Each column of $[\bar{R}^*]$ represents a pure, real target vector. These targets are the real factors sought in target factor analysis.

One such model data matrix, involving 16 rows and 5 columns, was constructed from 2 factors. The root-mean-square (RMS) of the error matrix, which was added to this pure data matrix, was 0.04. When the impure data matrix was subjected to abstract factor analysis, the resulting IE and IND functions [9] clearly indicated that 2 factors were responsible for the data and that the RE was 0.04168, in complete agreement with the known facts. A series of target vectors was constructed by adding various errors to the columns of $[\bar{R}^*]$. The detailed results of testing two such impure targets, labelled B and C, are shown in Table 1. Both of these targets contain different amounts of error and are meant to represent row target vectors for the first column, $\bar{R}_1$, of the row matrix. Also included in Table 1 is the pure target, $\bar{R}_1^*$. As the pure target is known, the various RMS errors can be calculated directly. These errors are shown near the bottom of the table. The theoretical errors calculated by means of eqns. (30), (33) and (34) are also listed in Table 1. The agreement between the known RMS error and that calculated from theory is excellent. In target-testing of real chemical data, pure data are not accessible

TABLE 1

Results of target testing a 16 × 5 model set of impure data containing two factors

$\bar{R}^*$ Pure	Target B			Target C		
	$\bar{\bar{R}}$ Impure	$\bar{R}$ Predicted	$\bar{R} - \bar{\bar{R}}$ Diff.	$\bar{\bar{R}}$ Impure	$\bar{R}$ Predicted	$\bar{R} - \bar{\bar{R}}$ Diff.
0	0.05	0.091	0.041	0.09	0.104	0.014
5	4.96	4.969	0.009	5.00	4.985	−0.015
10	9.99	10.051	0.052	10.18	10.067	−0.113
50	50.00	49.956	−0.044	49.97	49.878	−0.092
60	60.03	60.049	0.019	59.91	60.078	0.168
65	65.01	64.907	−0.103	64.97	64.931	−0.039
70	70.01	69.900	−0.110	69.91	69.924	0.014
74	74.01	73.899	−0.111	74.09	73.927	−0.163
80	80.05	79.962	−0.088	80.12	79.988	−0.132
102	101.99	102.023	0.033	102.06	102.052	0.008
105	104.93	105.071	0.141	105.09	105.100	0.010
110	109.94	110.014	0.074	110.09	110.066	−0.024
126	125.95	125.982	0.032	126.21	126.017	−0.193
135	135.00	135.049	0.049	135.00	135.088	0.088
140	140.03	139.949	−0.081	139.82	139.997	0.177
162	161.95	161.977	0.027	162.00	162.029	0.029
Known RMS error	0.037	0.060	0.074	0.105	0.063	0.102
RMS error from theory	0.048 RET	0.057 REP	0.074 AET	0.085 RET	0.056 REP	0.102 AET

TABLE 2

Summary of target testing a 16 × 5 model set of impure data generated from two factors

$\bar{R}_1$ Targets	AET	REP	RET	True (RET)	RELI	SPOIL
A	0.059	0.057	0.017	0	0.96	0.30
B	0.074	0.057	0.048	0.037	0.89	0.84
C	0.102	0.056	0.084	0.105	1.18	1.50
D	4.084	0.100	4.082	4.458	1.09	40.99
$\bar{R}_2$ Targets						
E	0.0143	0.0137	0.004	0	0.96	0.30
F	0.046	0.014	0.044	0.048	1.08	3.21
G	0.111	0.014	0.111	0.120	1.08	7.98
H[a]	0.077	0.0007	0.077	—	—	110.38

[a]This test vector was made of random numbers.

and the RMS errors cannot be calculated directly as done in the model studies. However, eqns. (30), (33) and (34), derived from the theory, allow these errors to be estimated.

Such target testing of the model data is summarized in Table 2, which includes the results of targets B and C just described. In calculating the RELI for the $\bar{R}_1$ targets and $\bar{R}_2$ targets, the true RET was used as the estimate of the real error in the target. Even when the error in the target was increased to 4.458, which is forty times greater than the error contributed by the data matrix, the RELI still remains close to 100%. The SPOIL values give an excellent indication of how much error would be introduced into the data matrix reproduction if these targets were employed. Targets A, B and C are acceptable but target D is obviously not an acceptable representation for $\bar{R}_1$, the first factor. Furthermore, target E is acceptable, target F is moderately acceptable, but targets G and H are not acceptable representations for $\bar{R}_2$, the second factor. Target H was made from random numbers, in order to examine how nonsensical targets behaved and if they would be rejected by the theory. Its extremely large SPOIL, 110.39, gives unquestionable evidence that it is not a factor.

The SPOIL values for targets A, B and E are less than unity; these targets should therefore lead to improvement in data matrix reproduction. However, the predicted targets will be less accurate than the test targets (e.g. target B in Table 1). Conversely, the SPOIL values for targets C, D, F and G are greater than unity. These targets are identified as being true factors because their RELI values are close to 100%, but unfortunately, utilization of these targets will lead to poor data matrix reproduction, the degree of spoilage being indicated by the magnitude of the SPOIL. There is one fringe benefit in this situation. Since these targets are true factors, the predicted targets will be more accurate than the original impure targets (e.g., target C in Table 1). In other words, insertion of a correct but highly inaccurate target into the target-testing scheme can lead to a predicted target which is more accurate than the test target if the error from the data matrix is less than the error in the test target.

APPLICATIONS

Mass spectra

The new theory has a wide variety of chemical applications. It is particularly useful for identifying components in mixtures. The mass spectral data of Ritter et al. [14] concerning two different data matrices can serve for illustration. The first data matrix was generated by measuring the intensities of 7 different mixtures of cyclohexane and hexane at 17 m/z positions. The second data matrix was generated by measuring the intensities of 4 different mixtures of cyclohexane and cyclohexene at 20 m/z positions. Ritter et al. showed that the rank of each of these two matrices was 2, giving clear evidence that two components were present. Use of the IE and IND functions led to the same conclusion [9], and showed that the real errors (RE) in the two data matrices were 0.13 and 0.15, respectively. Malinowski and McCue [15] showe in detail how target transformation could be used not only to identify but also to quantify the components in these mixtures.

When these two data matrices were subjected to target factor analysis, the results summarized in Table 3 were obtained. The target vectors consisted of the mass spectral intensities of the pure components measured at m/z position consistent with the data matrix. The RELI values indicate that cyclohexane is present in both sets of mixtures and that hexane is present in the first mixtu but not in the second. The presence of hexane in the second mixture is ruled out because the hexane target has a RELI of only 4% and a SPOIL of 77.76, which is excessively large. For the true components, the REP is almost identic to RET; this is expected because all data, including the target vectors, were measured by the same investigators under the same experimental conditions. Hence the real error in the data matrix should be identical to the RET and, consequently, identical to the REP.

From this study, it can be seen how the RELI and SPOIL functions are valuable in identifying the true components of a mixture and rejecting a bogus component. The methodology is not limited to mass spectral data but can be

TABLE 3

Target testing of mass spectral intensities of mixtures of cyclohexane/hexane and cyclohexane/cyclohexene[a]

Mixture	Target	AET	REP	RET	$(RET)_{est}$[b]	RELI	SPOIL
C_6H_{10}/C_6H_{12}	Cyclohexane	0.18	0.13	0.13	0.13	1.02	0.70
	Hexane	0.17	0.14	0.09	0.13	1.12	0.56
C_6H_{10}/C_6H_8	Cyclohexane	0.24	0.15	0.20	0.15	0.85	1.39
	Hexane	3.68	0.05	3.68	0.15	0.04	77.76

[a]Data taken from the work of Ritter et al. [14].

[b]Estimated from the real error values obtained from abstract factor analysis of the original data matrices (see [9]).

applied to other spectroscopic information such as absorption and emission spectra where the intensities of the lines (or some function of the intensities) can be expressed as a linear sum of the composition.

Nuclear magnetic resonance

The proposed theory has many possible applications in nuclear magnetic resonance (n.m.r.). One application concerns the study of solvent effects. Preliminary factor analysis studies [9, 11, 16] have shown that the chemical shift of a solute in a given solvent is a linear sum of contributing factors. By confining attention to simple, rigid, non-polar solute molecules, these investigations showed that, after correction for bulk-magnetic susceptibility of the solvent, three important factors are responsible for n.m.r. solvent shifts. From theoretical considerations, these factors are suspected to be: (1) the shift of the free solute molecule, which can be represented by the gas-phase shift of the solute, (2) van der Waals' interaction between solute and solvent molecules, and (3) solvent anisotropy. For example, an abstract factor analysis study [9] of the IE and IND functions of ^{19}F solvent shifts [17] clearly indicated that there are three factors and that the real error in the data is 0.035 ppm. Proving whether or not the gas-phase shift of the solute is a real factor in this situation is an ideal problem for target factor analysis. Of the 19 solutes employed in the study, only 12 were subjected to vapor-shift determinations because of experimental difficulties in making such measurements. These 12 values, (Table 4) were used to construct the test vector. The predicted shifts which emerged from target factor analysis are also given in Table 4. An advantage of target testing is that the missing data points are predicted. From the differences between the test and predicted points, AET can be obtained from eqn. (30), which is applicable even though the test vector is incomplete. The values for REP and RET, as well as AET, are listed at the bottom of Table 4.

TABLE 4

Target testing ^{19}F gas-phase shifts

Solute	Test Vector $\bar{R}$	Predicted vector $\bar{R}$	Diff. $\bar{R} - \bar{\bar{R}}$	Solute	Test vector $\bar{R}$	Predicted vector $\bar{R}$	Diff. $\bar{R} - \bar{\bar{R}}$
CF_2Br_2	−2.27	−2.50	−0.23	CF_2CCl_2	95.67	95.98	0.31
$CFCl_3$		4.98		CF_3CCF_3	59.29	59.41	0.12
CF_2ClBr	4.64	4.52	−0.12	C_4F_8	140.85	140.84	−0.01
$CFCl_2CFCl_2$		71.62		CF_4		68.70	
s-$C_6F_3Cl_3$		118.98		$C_6H_5CF_3$		69.94	
CF_2Cl_2	12.17	11.95	−0.22	$CF_3CHClBr$		82.66	
Cis - CFClCFCl	111.91	112.14	0.23	α-C_6F_{14}	86.33	86.22	−0.11
Trans - CFClCFCl		125.62		β-C_6F_{14}	130.37	130.26	−0.11
C_6F_6	170.73	170.77	0.04	γ-C_6F_{14}	126.73	126.44	−0.29
CF_3CCl_3	87.04	86.97	−0.07				
				THEORETICAL ERRORS	0.19 RET	0.05 REP	0.19 AET

The SPOIL is calculated to be 3.8, which means that if this target is used for data matrix reproduction, considerable error will be introduced. Essentiall the experimental error in the vapor measurements is almost four times the error in the solution measurements. This agrees with the well-known fact that vapor measurements are much more difficult to carry out. From experimental information [17], the error in the vapor shifts can be estimated as approximately 0.14, which gives a RELI value of approximately 80%; in so far as RELI exceeds 50%, the gas-phase shift must be a real factor. Most importantly, this can be concluded independently of the other two participating factors.

Because REP is considerably smaller than RET, it can be concluded that the predicted shifts in Table 4 emulate the true shifts better than the raw test vector. This is analogous to the model study involving target C (see Table 1). Thus, target factor analysis can be useful not only for determining whether or not a given target is a true factor but also for purifying the raw target data.

The author thanks Harry Rozyn and Phuong T. Dang, both supported by the College Work Study Program, for their assistance with the computer program and computations.

REFERENCES

1 B. R. Kowalski, (Ed.), Chemometrics: Theory and Applications, ACS Symposium Series 52, American Chemical Society, Washington, DC, 1977.
2 D. G. Howery, Am. Lab., 8 (2) (1976) 14.
3 P. H. Weiner, Chem. Tech., May (1977) 321.
4 R. J. Rummel, Applied Factor Analysis, Northwestern Univ. Press, Evanston, Ill., 1970.
5 B. Fruchter, Introduction to Factor Analysis, Van Nostrand, Princeton. N.J., 1954.
6 A. L. Comrey, A First Course in Factor Analysis, Academic Press, New York, 1973.
7 D. N. Lawley and A. E. Maxwell, Factor Analysis as a Statistical Method, American Elsevier, New York, 1971.
8 P. Horst, Factor Analysis of Data Matrices, Holt, Rinehart and Winston, New York, 196
9 E. R. Malinowski, Anal. Chem., 49 (1977) 606, 612.
10 E. R. Malinowski, Ph.D. Thesis, Stevens Institute of Technology, 1961; Dissertation Abstracts, 23 (8) (1963) 62-2027.
11 P. H. Weiner, E. R. Malinowski and A. R. Levinstone, J. Phys. Chem., 74 (1970) 4537.
12 E. R. Malinowski, D. G. Howery, P. H. Weiner, J. M. Soroka, P. T. Funke, R. B. Selzer and A. Levinstone, FACTANAL-Target-Transformation Factor Analysis, Program 320, Quant. Chem. Prog. Exch., Indiana University, Bloomington, IN, 1976.
13 D. L. Duewer, A. M. Harper, J. R. Koskinen, J. L. Fasching and B. R. Kowalski, ARTHUR, Version 3-7-77.
14 G. L. Ritter, S. R. Lowry, T. L. Isenhour and C. L. Wilkins, Anal. Chem., 48 (1976) 591
15 E. R. Malinowski and M. McCue, Anal. Chem., 49 (1977) 284.
16 P. H. Weiner and E. R. Malinowski, J. Phys. Chem., 75 (1971) 1207, 3160.
17 R. J. Abraham, D. F. Wileman and G. R.Bedford, J. Chem. Soc., Perkin Trans. 2, (1973) 1027.

Analytica Chimica Acta, 103 (1978) 355—365
Computer Techniques and Optimization

HOSE — A NOVEL SUBSTRUCTURE CODE*

W. BREMSER
Hauptlaboratorium, BASF AG, D-6700 Ludwigshafen (West Germany)
(Received 3rd May 1977)

SUMMARY

A novel system of substructure codes has been developed to characterize the spherical environment of single atoms and complete ring systems. The codes are generated automatically from topologically represented chemical structures and serve to describe structural entities corresponding to spectral parameters uniquely. Their hierarchical order permits desired substructures and the corresponding chemical shifts to be sought in inverted files generated from a larger data base, thereby facilitating the estimation of unknown spectra.

The need for structure elucidation in modern organic chemistry can only be satisfied by fast and reliable spectroscopic techniques. Consultation of detailed and well-arranged reference material with a precise answer to specific questions is necessary in order to eradicate dangerous misinterpretations as well as eliminate the costly synthesis and recording of reference compounds which often are not readily available. The interaction between structural and spectroscopic properties, however, can only be accessed in computer search algorithms or represented in registers when the link between structure and spectrum has already been established in the data base, as has been demonstrated for the field of computer-aided interpretation of ^{13}C-n.m.r. spectra [1, 2].

It is commonly accepted that the ^{13}C-n.m.r. resonance frequencies are influenced by their nearest neighbours. The well-known α-, β-, γ- and δ-effects [3] and their decreasing magnitude lead to the concept of linking substructures to subspectra [1] and even single atoms of specified environment to single resonances. The latter concept required substructure codes representing the spherical arrangement of substituents.

Besides the linear notation developed by Wiswesser [4], various other codes have been designed to characterize the environment of individual atoms or functional groups e.g. the group code [5, 6] to characterize n.m.r. spectra. Both the idea of augmented atoms by Lynch et al. [7] and the DARC system of Dubois [8] describe substructures contained in the molecule. Finally the CIDS Chemical Search Keys [9] allow the identification of functional groups with the help of a numerical code.

---

*Reprinted in part in the introduction of [10].

The substructure codes proposed here were designed with four objectives in mind: full compatibility with the topological representation without predefining specific fragments; description of at least three, preferably four, spheres with low storage requirements; hierarchically ordered, yet mnemonically easy to understand codes; unique and automatic code generation from the topological matrix [10].

SPHERICAL DESCRIPTION OF ENVIRONMENT

The idea of describing the environment of individual atoms spherically has been explained elsewhere [1]. However, experience with its rigid format and incomplete set of symbols prompted the development of a new coding system with the priority rules defined in Table 1. Now the multiple bonds characterizing the hybridization and thereby the range of chemical shifts have highest priority. The number of bits of each symbol that define the numerical value of the code, and thus its position in the hierarchical register, varies with its probability from 3 to 12. This guarantees that for common structural elements with the symbols "C", "*" and the sphere separators "," or "(" and "/" enough bits are left in the two available computer words (72 bits) to characterize up to four spheres, even if the molecule is relatively branched. For odd structural elements with heteroatoms and lower probability, the code breaks off at the end of the second storage word, normally in the middle of the third or fourth sphere.

In Fig. 1, the slightly modified example from earlier work [1] is represented by the following Hierarchically Ordered Spherical Description of Environment (HOSE) code: the first sphere (substituents of carbon 1) contains three atoms:

2a	*C	*C*CC(
2b	*C	656550
2c	C	

2a has higher priority than 2b, because its substituent 3a is more highly substituted than 3b. The bracket symbolizes the end of the first sphere. The octal code is given for comparison. The substituents of these atoms are then listed in the predefined sequence and separated by commas:

2a	→	3a	*C	*C,*C,=OC/
2b	→	3b	*C	6516517450
2c	→	3c	=O	
		3d	C	

The sphere separator (/) between the second and the third sphere marks the end of the second sphere. The code for the third sphere is composed as follows:

3a	→	4a	*C	*CX,*&,,CC/
		4b	X	6531162411550
3b	→	4a	*&	
3c	→	none		
3d	→	4c	C	
		4d	C	

TABLE 1

List of symbols in decreasing priority for substructure codes (HOSE). It should be noted that symbols for delocalized charges ('—') and for localized charges greater than 1 ('+4') are enclosed in single quotes. MM signify symbols of elements no. 1 through no. 63 and NN symbols of heavy elements starting from 64.

Symbol	Meaning	Octal	No. of bits
R	ring	77	6
%	≡	76	6
=	=	7	3
*	...	6	3
C	C	5	3
Ø	O	4	3
N	N	37	6
S	S	36	6
P	P	35	6
Q	Si	34	6
B	B	33	6
F	F	32	6
X	Cl	31	6
Y	Br	30	6
;	truncation	3?	3
I	I	27	6
'NN'	NN	26YY	12
'MM'	MM	25XX	12
#	truncation	25? ?	6
&	ring closure	24	6
+	positive charge	23	6
—	negative charge	22	6
>	coordinate bond	21	6
D	D	20	6
:	truncation	2?	3
,	separator	1	3
(//)	sphere separators	0	3
;	sphere separator after heterofocus	none	0

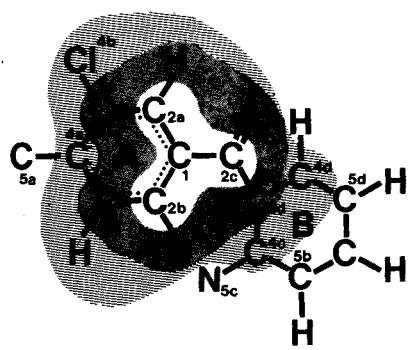


Fig. 1. The four spheres surrounding the focus C_1 with the hierarchical numbering defined by the priority rules in Table 1.

4c has higher priority than 4d because of its higher degree of substitution. 4a has already been coded as a substituent of 3a, therefore the bond between 3b and 4a represents a ring closure. Because the HOSE code breaks off at the end of the second computer word (72 bits), the listed code will read

```
*C*CC(*C,*C,=OC/*CX,*&
6565506516517450653116 24
```

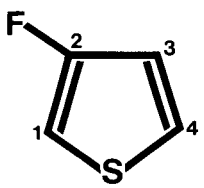
In this example of a highly substituted carbon, only two and a half spheres are defined by the substructure code. For completeness, the code for the fourth sphere is

4a	→	3b	*&	*&C,,CN,C)
		5a	C	624511537150
4b	→	none		
4c	→	5b	C	
		5c	N	
4d	→	5d	C	

It should be remembered that the substituents of 4a are calculated only once; the ring closure sign automatically terminates a coding pathway. Similarly, bonds to delocalized charges or coordinate bonds to π-electron systems are coded only in one direction and not treated as a ring structure (cf. HORD codes).

If the central atom of a substructure ("focus") is a different element than carbon, its element symbol precedes the code of the first sphere. One should bear in mind that this case is numerically characterized by a zero followed by the atomic number (octal). Therefore the hierarchical sequence is defined by the periodic table (heavy elements preceding light elements). An example for a code focused on a heteroatom is the sulphur in 3-fluorothiophene (I):

```
S;CC(=C,=C/F&,&/)
02055075175032241240 0000
```



(I)

RING STRUCTURES

In many cases in chemistry larger molecular entities ("superatoms") are more interesting than the surroundings of a single atom. These preformed structural units will be mostly ring systems and therefore the original concept was modified to incorporate a similar code for Hierarchically Ordered Ring Description (HORD). The meanings of the symbols and their priority are defined in Table 2.

A ring in the sense of the HORD codes is a group of atoms connected by a cyclic pathway up to 9 atoms in length; 10-membered rings are regarded as

TABLE 2

List of symbols in decreasing priority for ring codes (HORD)

Group	Symbol	Meaning	Octal	No. of bits
1	R9	9-ring	777	9
1	R8	8-ring	776	9
1	R7	7-ring	775	9
1	R6	6-ring	774	9
1	R5	5-ring	773	9
1	R4	4-ring	772	9
1	R3	3-ring	771	9
1	R0	no ring	770	9
2	%<	≡ and ≡ ,=, ···	7	3
2	%	≡	6	3
2	==	= and =	5	3
2	=*	= and ···	4	3
2	=	=	3	3
2	#	aromatic	2	3
2	*	···	1	3
2		saturated	0	3
3	Ø	O	7	3
3	N	N	6	3
3	S	S	57	6
3	P	P	56	6
3	Q	Si	55	6
3	B	B	54	6
3	F	F	53	6
3	X	Cl	52	6
3	Y	Br	51	6
3	I	I	50	6
3	'NN'	NN	4YY	9
3	'MM'	MM	3XX	9
3	.	C intermediate	2	3
4	&	bridgeheads follow	1	3
5	(	begin 2nd sphere	0	3
5	HOSE codes of substituents separated by ","			

open-chain compounds as they show only a small amount of distortion and restricted mobility. In bi-, tri- and polycyclic compounds, all possible rings with up to 9 members will be described by HORD codes. However, members of a ring can only be connected by classical bonds; this excludes cyclic structures linked by coordinate bonds e.g. chelates, sandwich compounds, etc. HORD codes can be distinguished from HOSE codes by the preceding R. The focus now does not reside on a single atom, but in the middle of the ring. Therefore the ring members form the first sphere, their substituents the second sphere. The ring itself is characterized by ring size (group 1), bond type (group 2), elemental composition (group 3) and bridge heads (group 4 of symbols). The second sphere is described by the HOSE code symbols defined in Table 1 with the comma separating the substituents of the ring members.

The procedure is exemplified by the two rings in Fig. 1. Although both ring are six-membered, ring A is aromatic and ring B is saturated. As there are neither heteroatoms nor bridgeheads present, the codes for the first sphere read (including octal representation):

ring A	R6 # (	77420
ring B	R6(	77400

The numbering is defined in these two cases by the priority of the substituents. For ring B, carbon substituent 2c has higher priority than nitrogen 5c. Thus the numbering will start at 3d and go the shortest way to the next highest substituent (here 5c, therefore counterclockwise). In ring A, carbon substituent 5a and 2c are equivalent but 4b is nearer to 5a and therefore a clockwise numbering starting with 4a will give highest priority:

ring A	R6 # (C,X,,C,,)	774205131115110000000000
ring B	R6(C,N,,,,)	774005137111100000000000

In 3-fluorothiophene (I), the numbering starts with the highest valent bond and with shortest pathway to the next highest bond. Because the double bonds C_1=C_2 and C_3=C_4 have equal priority, a start at C_1 with clockwise rotation and at C_4 with counterclockwise rotation would result in ambiguous numbering. However, the fluorine substituent in the second sphere gives the first solution a higher numerical value of the HORD code:

R5==....S(,F,,,)	773522225701321110000000

If, however, bond C_3—C_4 were a single bond, the heteroatom sulphur would cause a counterclockwise numbering beginning at C_2:

R5=..S(F,,,,)	773322570321111000000000

It should be noted that the two additional symbols for intermediate C are omitted and that the numerical value for selected heteroatoms differs for the ring atoms in HORD codes and for the substituents in HOSE (or HORD) codes. Consultation of Tables 1 and 2 and some practical experience will provide the knowledge necessary to read and understand the description of structural environment of both atoms and rings. Furthermore a series of practical hints [10] will help the beginner to become familiar with the hierarchical coding system and enable the desired canonical code, as entered in the registers, to be found.

APPLICATIONS OF HOSE/HORD SUBSTRUCTURAL CODES

The possible applications of these substructure codes are manifold. In this paper a survey of five typical aspects connected with ^{13}C-n.m.r. spectroscopy is given.

Characterization of structural elements

When the different entries in an inverted file are scanned, the decision whether a reference compound characterized by a number, a chemical name,

or a molecular formula represents a suitable model compound can often be made only by looking up the chemical structure of the molecule in question. Therefore a short descriptive code defining the structural environment of a particular atom will give an intuitive survey of its hybridization, substitution and charge. Therefore the HOSE codes included in the registers of chemical shifts, coupling constants and relaxation times [10] save the effort of looking up unsuitable reference compounds in many cases.

Substructural registers

Inverting the file to decreasing arithmetical values of HOSE/HORD-codes provides a substructure index which is probably the most important tool for the practising spectroscopist. It leads the way from the concept of a certain substructure to all carbon atoms with similar environment regardless of the rest of the molecule [1]. Therefore the substructure index should list the HOSE codes of all atoms (except ^1H which is omitted in the Morgan representation [11]) in all molecules contained in a data collection. The HORD codes of all possible ring structures are calculated similarly. Following the hierarchical concept, HORD codes are listed first followed by the HOSE codes of carbon atoms and the HOSE codes of non-carbon atoms (all with decreasing numerical value of the code). Behind the code all compounds containing the desired substructure are listed with their reference and atom number. The entry R0 between HORD and HOSE contains the references without any ring structure. An example is given in Table 3.

Computer search

The search for substructures or combinations of substructures can also be done with the help of an interactive programme [12]. Here, pointing to a specific node in the molecule saves the effort of writing down the hierarchical code. The actual procedure, however, is identical. Now the computer generates the code, looks up the entry in the inverted file (i.e. in the register) and presents the result [13]. The superiority of the computer is based on its ability to combine the results of several search runs by logical AND, OR, or NOT. The disadvantage lies in its high cost, low portability and the exclusion of an intuitive survey of neighbouring entries in the register.

Chemical shift ranges

The interpretation of the spectrum of an unknown compound can be divided into two different steps which are normally passed several times in an iterative process until the structure elucidation can be regarded as complete. The first step is the postulation of structural elements explaining the spectral features observed during the experiment. The second step is the reverse process, i.e. the answer to the question: "How would the spectrum look, if the postulated structure were correct?". This structure verification should be based on a precise knowledge of the effects of substitution on spectral parameters. Therefore numerous refined systems of substituent increments [14] have

TABLE 3

Segment from the HOSE code register leading to the individual references (characterized by compound number and atom number) in the data collection [10].

CCS(C,,C/=OC,*C*&/,*C*	9927–01			
CCS(C,,C/CO,C&/,&,)	8324–02			
CCS(C,,C/C,&/&)	4676–01			
CCS(C,,C/O,CC&/&,,)	8321–02			
CCS(C,,C/O,C&/&,)	8322–02			
CCS(O,,C/&,CC&/,)	8320–02			
CCS(O,,C/&,&/)	8316–02			
CCS(S,,C/&,&/)	8315–02			
CCS(,,=O=OX/,,/)	9334–02			
CCS(,,=OX/,/)	9292–01			
CCS(,,CC+/,/)	2986–01			
CCS(,,C/=NX/C,)	6661–08			
CCS(,,C/CO/C,C)	352–06			
CCS(,,C/NN+/CC,CC)	1362–06			
CCP'W' (=CO,=CO,CCC+,:	9456–01			
CCP'W' (=CO,=CS,CCC+,	9454–01			
CCP'W' (=CO,*C*C,CCC+,	9453–01			
CCP'W' (=CS,=CS,CCC+	9455–01			
CCP'W' (=CS,*C*C,CCC+	9452–01			
CCP'W' (*C*C,*C*C,CCC+	9451–01			
CCP–(=OO,,=OOO/,C,,C,C	8471–01			
CCP(=CC,C,=OCC/Y&,C,:	3646–08	3645–08		
CCP(=CC,C,=OCC/&,C,&,	3643–08	3642–08		
CCP(=CC,C,=SCC/&,C,&	4306–08	4305–08		
CCP(=CC,C,CC/&,C,&,&	3648–08	3647–08		
CCP(=OO,,=OOO/,C,,C,C/,	8469–01			
CCP(CC,C,=OCC/CC,C,C&,	4506–01	4505–01		
CCP(CC,C,CC/CC,C,C&,*C	4504–01	4503–01		
CCP(C,C,=OCC/=OC,C,,*C*	4469–05	4469–01	4468–05	4468–01
CCP(C,C,=OCC/CO,C,,*C*C	4470–05	4470–01		
CCP(C,C,=OCC/C,C,,*C*C,	4467–05	4467–01		
CCP(C,C,=SCC/C,&,,,/C	6405–01	6404–01	6403–01	6402–01
CCP(C,C,=SCC/C,&,,,/:	6401–01			
CCP(C,C,CCC+/C,C,*C*C,	3862–19			
CCP(C,C,CCC+/C,&,*C*C	3861–19			
CCP(C,C,CCC+/C,&,,,/C	6410–01	6409–01	6408–01	6407–01
CCP(C,C,CCC+/C,&,,,/:	6406–01			
CCP(C,C,CCC+/&,&,*C*	3860–19			
CCP(C,C,CC/C,&,,/C&)	6394–01	6393–01	6392–01	6391–01
CCP(C,C,CC/C,&,,/&)	6390–01			
CCP(C,C,OO/C,&,C,C/C&	6400–01	6399–01		
CCP(C,C,OO/C,&,C,C/&,	6398–01			
CCP(C,C,XX/C,&,,/C&	6397–01	6396–01		
CCP(C,C,XX/C,&,,/&,	6395–01			
CCP(C,C,/C,&/C&)	6389–01	6388–01		
CCP(C,C,/C,&/&)	6387–01			
CCP(C,&,CCC+/&,*C*C,	3859–19			
CCP(&,&,=CCC/,,/)	9527–01			
CCP(&,&,CCC+/*C*C,*C	3858–19			

TABLE 3 (continued)

CCP(,,=OCC/,*C*C,*C*C/*	2802–13				
CCP(,,=SCC/,C,C/C,&)	4271–06				
CCP(,,CCC+/*C*C,*C*C,*	3856–19				
CCP(,,CC/C,C/C,&)	4266–06				
CCQ(=CC,*C*C,CCC/&,,*C	1474–01				
CCQ(=C,*C*C,CCC/&,*C&	1472–01				
CCQ(C,C,CCC/C,&,,,/&,	4484–01				
CCB(=OO,CN,>N/,C,,C:	9594–02	9593–02	9590–02	9589–02	9588–02
CCB(CC,,CC/,,=C,CC/C,CC	3262–12	3262–07			
CCFF(=CO,CFF,,/C,–,	4763–03				
CCFF(=CF,CFF,,/FF	6418–03				
CCFF(=CX,CFF,,/CX,	4334–06	4334–05			

been developed to allow the prediction of, e.g., chemical shifts of specific substructures. However, they are normally restricted to well-defined classes of compounds and a limited number of substituents. Interactions between the individual substituents are often not taken into account.

The HOSE codes allow a much more detailed description of substructures. The register depicted in Table 4 contains the averaged chemical shifts with standard deviation, the number of resonances for this HOSE code contained in the data collection [10], and the highest and lowest chemical shifts encountered. A modified average is calculated after eliminating 1/6 of the chemical shifts (those with the highest deviation from the mean). These entries in the register give typical ranges of chemical shifts expected for certain structural elements. Furthermore, the hierarchical ordering often allows the estimation of chemical shifts from the neighbouring entries when the precise substructure is not contained in the register. Expected chemical shifts can also be derived for ranges of substructure codes with identical first, second or third sphere, and strong deviations point to specific interactions between the substituents or errors in the data collection which can be automatically detected this way.

Automatic estimation of unknown spectra

As shown for the substructure search, this procedure can also be computerized fully in order to estimate the ^{13}C-n.m.r. spectrum of a postulated chemical structure automatically [15]. An interactive program [13] allows building and visualizing the desired compound, generation of the topological matrix and calculation of the HOSE codes. Finally, the desired entries are looked up in the register and the expected chemical shift ranges for the individual carbon atoms are tabulated, including standard deviation and number of entries.

A program of this type is an important step towards the automation of structure elucidation. It will be supported by programs for the identification of substructures [1, 2] and their combination with the aim of presenting suggestions for possible structural entities.

TABLE 4

Segment from the HOSE code register with averaged chemical shifts, standard deviations, number of entries and highest/lowest values in the data collection [10].

=CCC(C,=CC,CC/C,C,CC&,C	136.0 # 0.4 PPM	2 LINES (136.3–135.7 PPM)
=CCC(C,=CC,CC/O,O,=O&,=	125.2 # 0. PPM	1 LINES (125.2–125.2 PPM)
=CCC(C,=CC,O/=O&,C,CCC,	173.1 # 0.5 PPM	2 LINES (173.5–172.8 PPM)
=CCC(C,=CC,/=OC,C,CC&/,	153.5 # 0. PPM	1 LINES (153.5–153.5 PPM)
=CCC(C,=CC,/=OC,&,/,CC:	155.0 # 0. PPM	1 LINES (155.0–155.0 PPM)
=CCC(C,=CC,/=OO,C&,=OO,	154.4 # 0. PPM	1 LINES (154.4–154.4 PPM)
=CCC(C,=CC,/=SO,C&,=OO	146.8 # 0. PPM	1 LINES (146.8–146.8 PPM)
=CCC(C,=CC,/CCC,&,/=O&	127.8 # 0. PPM	1 LINES (127.8–127.8 PPM)
=CCC(C,=CC,/C,C,CC&/&,	134.1 # 0. PPM	1 LINES (134.1–134.1 PPM)
=CCC(C,=CO,/=OC,CC,C/,,=	144.0 # 0.2 PPM	3 LINES (144.2–143.9 PPM)
=CCC(C,=CO,/=O&,C,&/,=	153.0 # 0.9 PPM	2 LINES (153.7–152.4 PPM)
=CCC(C,=CO,/CCO,CC,C/CO,	130.8 # 1.7 PPM	10 LINES (134.0–127.4 PPM)
=CCC(C,=CO,/CC,CC,C/CO,,	127.9 # 0. PPM	1 LINES (127.9–127.9 PPM)
=CCC(C,=CN,/C,CC,C&/&	129.9 # 0. PPM	1 LINES (129.9–129.9 PPM)
=CCC(C,=CX,=CX/CC,X&	131.6 # 0. PPM	1 LINES (131.6–131.6 PPM)
=CCC(C,=CX,=ON/XXX,	128.9 # 1.3 PPM	2 LINES (129.8–128.0 PPM)
=CCC(C,=C&,C&/,CC,*C*C	128.2 # 0. PPM	2 LINES (128.2–128.2 PPM)
=CCC(C,=C>'FE',C/=OO,C,	107.9 # 0. PPM	1 LINES (107.9–107.9 PPM)
=CCC(C,=C,=C/CCC,&,&/,	141.8 # 0. PPM	1 LINES (141.8–141.8 PPM)
=CCC(C,=C,=C/C,&,&/C)	146.0 # 0. PPM	1 LINES (146.0–146.0 PPM)
=CCC(C,=C,=OC/=OC,C,,=&	148.5 # 0. PPM	1 LINES (148.5–148.5 PPM)
=CCC(C,=C,*C*C/=CC,CO,*C	136.5 # 3.0 PPM	4 LINES (140.9–134.8 PPM)
=CCC(C,=C,CCC/=OC,C&,,,	162.7 # 0. PPM	1 LINES (162.7–162.7 PPM)
=CCC(C,=C,CCC/=OC,C,=&,	162.6 # 0. PPM	1 LINES (162.6–162.6 PPM)
=CCC(C,=C,CCC/=OC,C,C&,	160.6 # 2.2 PPM	7 LINES (163.2–157.2 PPM)
=CCC(C,=C,CCC/=OO,C,C&,	161.3 # 0. PPM	1 LINES (161.3–161.3 PPM)
=CCC(C,=C,CCC/C,C,C&,&	141.4 # 0. PPM	1 LINES (141.4–141.4 PPM)
=CCC(C,=C,CCO/=C,&,=CC,	145.5 # 1.6 PPM	2 LINES (146.6–144.4 PPM)
=CCC(C,=C,CC/*C*C,O,,/*C	145.2 # 0. PPM	1 LINES (145.2–145.2 PPM)
=CCC(C,=C,CN/C,C,CC&,C	134.8 # 0.8 PPM	2 LINES (135.3–134.2 PPM)
=CCC(C,=C,C/=C,C,/C,=CC,	147.3 # 0.7 PPM	3 LINES (148.0–146.7 PPM)
=CCC(C,=C,/%C,C/C,=C)	147.2 # 1.5 PPM	11 LINES (150.4–146.5 PPM)
=CCC(C,=C,/%C,C/C,O)	145.2 # 0. PPM	1 LINES (145.2–145.2 PPM)
=CCC(C,=C,/=C,=C/C,CC)	133.8 # 0.4 PPM	6 LINES (134.0–133.3 PPM)
=CCC(C,=C,/=C,C/CO,=CC)	142.5 # 0.8 PPM	2 LINES (143.0–141.9 PPM)
=CCC(C,=C,/=C,C/C,=C)	137.6 # 3.8 PPM	57 LINES (158.8–134.9 PPM)
=CCC(C,=C,/=C,C/C,=OC)	133.9 # 0.5 PPM	2 LINES (134.2–133.5 PPM)
=CCC(C,=C,/=C,C/C,*C*C)	138.6 # 1.6 PPM	4 LINES (140.7–136.9 PPM)
=CCC(C,=C,/=C,C/C,CCO)	137.4 # 2.3 PPM	3 LINES (139.4–134.9 PPM)
=CCC(C,=C,/=C,C/C,CC)	136.9 # 2.5 PPM	3 LINES (139.8–135.5 PPM)
=CCC(C,=C,/=C,O/C&,&)	136.7 # 0. PPM	1 LINES (136.7–136.7 PPM)
=CCC(C,=C,/=OC,&/,=O&,	152.0 # 0. PPM	1 LINES (152.0–152.0 PPM)
=CCC(C,=C,/=OC,&/,CC&,	153.3 # 0.6 PPM	2 LINES (153.7–152.8 PPM)
=CCC(C,=C,/=OO,C/,C,=C)	152.5 # 0.7 PPM	11 LINES (154.1–151.1 PPM)
=CCC(C,=C,/=OO,C/,C,CCO,	150.8 # 0. PPM	1 LINES (150.8–150.8 PPM)

Limitations

It should be mentioned in this context that the topological representation of chemical structures as used by CAS or IDC [11, 16] reflects only the molecular constitution, not the configuration or conformation. Therefore the present version of HOSE/HORD substructural codes also shows ambiguity in the representation of isomers, e.g. *cis—trans*, *exo—endo*, *syn—anti*, α—β, etc. The register of chemical shift ranges and the corresponding computer search program gives two average chemical shifts when two distinct maxima are detected in the distribution of resonance frequencies. Normally the experienced spectroscopist will find no difficulty in selecting the correct shift corresponding to the desired isomer.

I am indebted to B. Franke, H. Wagner, and E. Frank for help in designing structured inverted files and for calculating and arranging the depicted registers The BASF/GBF data collection (at present 13,000 spectra) was collected in cooperation with L. Ernst and C. Köhler.

REFERENCES

1 W. Bremser, M. Klier and E. Meyer, Org. Magn. Reson., 7 (1975) 97.
2 W. Bremser, Fresenius Z. Anal. Chem., 286 (1977) 1.
3 D. M. Grant and E. G. Paul, J. Am. Chem. Soc., 86 (1964) 2984.
4 W. J. Wiswesser, A Line-formula Chemical Notation, Crowell, New York, 1954.
5 N. S. Bhacca, L. F. Johnson and J. N. Shoolery, Varian NMR Spectra Catalog, Vol. 1 and 2, Varian Ass., Palo Alto, 1962 and 1963.
6 L. F. Johnson and W. C. Jankowski, Carbon-13 NMR Spectra, Wiley-Interscience, New York, 1972.
7 M. F. Lynch, J. M. Harrison, W. G. Town and J. E. Ash, Computer Handling of Structural Information, Elsevier, London, 1971.
8 J. E. Dubois, Israel J. Chem., 14 (1975) 17.
9 Handbook of CIDS Chemical Search Keys, Fein-Marquart Ass., Inc., Towson, 1973.
10 W. Bremser, L. Ernst and B. Franke, Carbon-13 NMR Spectral Data, Verlag Chemie, Weinheim, 1978.
11 H. L. Morgan, J. Chem. Doc., 5 (1965) 107.
12 Time share system, Honeywell Bull., 6050.
13 W. Bremser, to be published.
14 E. Pretsch, T. Clerc, J. Seibl and W. Simon, Tabellen zu Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden, Springer-Verlag, Berlin, 1976.
15 J. T. Clerc and H. Sommerauer, Anal. Chim. Acta, 95 (1977) 33.
16 E. Meyer, Angew. Chem., 15 (1970) 605.

Analytica Chimica Acta, 103 (1978) 367—378
Computer Techniques and Optimization

COMPUTERIZED KALOUSEK POLAROGRAPHY

M. BOS

Department of Chemical Technology, Twente University of Technology, Enschede (The Netherlands)

(Received 3rd January 1978)

SUMMARY

A versatile online computer-based system for Kalousek polarography features simultaneous recording of oxidation and reduction currents, averaging over successive scans, and processing of the polarographic data by curve-fitting. The accuracy for determinations of cadmium, potassium and lithium down to 10^{-5} M is ± 5% for pulse rates up to 25 Hz.

Kalousek polarography has received little attention in recent years [1], although Kinard et al. [2] outlined important analytical applications of the method, such as the ability to distinguish between reversible and irreversible systems and the possibility of determining reversible systems in the presence of proton discharge. The theoretical foundations of Kalousek polarography are well established. Kambara [3] was the first to solve the differential equations for the diffusion of compounds undergoing reversible oxidation/reduction during electrolysis with a square-wave voltage at a stationary plane electrode.

Koutecký [4] extended the theory to a dropping electrode; his general equations were modified by Ružić [1] to show the mean current versus potential relationships for the various types of Kalousek measurements.

Computerization of the Kalousek method was undertaken to make the technique available to general online computer owners who lack the complicated dedicated equipment for Kalousek polarography. Furthermore, this work was done to study the possibilities of automatic processing of the data, the accuracy that can be obtained by the computerized method, and the decrease in the detection limit obtained by averaging successive scans.

To be economically attractive, the method should work in a multi-user environment, for despite the decreasing cost of mini- and micro-computers themselves, mass storage devices and interfaces like A/D- and D/A-converters make up a large part of the total cost of a system. Only if these peripherals can be shared among a number of applications, e.g. titrations, coulometric and polarographic determinations, etc., can a general online computer system compete successfully with dedicated equipment for these techniques.

Additional advantages of replacing dedicated titrimetric and electroanalytical

equipment by an online computer are the great flexibility and the capability of data processing offered by the computer system.

Application of the theory of Kalousek polarography

Koutecký [4] derived the equation for the instantaneous current that flows when a square-wave potential with $E = E_1$ for $mT \leqslant t < (m + \frac{1}{2})T$, and $E = E_2$ for $(m + \frac{1}{2})T \leqslant t < (m + 1)T$ is applied to a dropping electrode where a reversible system undergoes reduction/oxidation:

$$i = nFq\ (7D_O/3\pi)^{\frac{1}{2}}\ \ t^{2/3}\{\mu_1 t^{-7/6} + \sum_{j=1}^{L}(\mu_1 - \mu_2)\,(-1)^j/[t^{7/3} - (jV)^{7/3}]^{\frac{1}{2}}\} \quad (1)$$

(The symbols used are listed in Table 1.) In this equation, $2V = T$ and $L = 2N$ for the case $NT < t < (N + \frac{1}{2})T$; and $2V = T$ and $L = 2N + 1$ for the case $(N + \frac{1}{2})T < t < (N + 1)T$. Also

$$\mu_1 = \left\{\exp\left[\frac{nF}{RT}(E_1 - E_0)\right]\ C_R^* - C_O^*\right\} \Big/ \left\{1 + (D_O/D_R)^{\frac{1}{2}} \exp\left[\frac{nF}{RT}(E_1 - E_0)\right]^{\frac{1}{2}}\right\}$$

$$\mu_2 = \left\{\exp\left[\frac{nF}{RT}(E_2 - E_0)\right]\ C_R^* - C_O^*\right\} \Big/ \left\{1 + (D_O/D_R)^{\frac{1}{2}} \exp\left[\frac{nF}{RT}(E_1 - E_0)\right]^{\frac{1}{2}}\right\}$$

Formula (1) can be rewritten to show separately the time-dependent and potential-dependent factors:

$$i = A\ \{F_1(t)\,\mu_1 + F_2(t)\,(\mu_1 - \mu_2)\} \quad (2)$$

with

$$A = nFq\ (7D_O/3\pi)^{\frac{1}{2}},\ F_1(t) = t^{-\frac{1}{2}}$$

and

$$F_2(t) = t^{2/3} \sum_{j=1}^{L} (-1)^j/[t^{7/3} - (jV)^{7/3}]^{\frac{1}{2}}$$

F_1 and F_2 were calculated for instantaneous current measurement at the center of the last half-period of E_1, as well as for the center of the last half-period of E_2, for various numbers of pulses per drop life of 960 ms. The results are given in Table 2.

For a fixed value of E_1, and for E_2 changing stepwise with time, eqn. (2) results in the familiar S-shaped polarogram also known in sampled d.c. polarography. Disregarding the sign of the current, the Kalousek polarographic data can be fitted to the equation

$$i' = i_l/\ \{\exp[(E_2 - E_{\frac{1}{2}})/S] + 1\} \quad (3)$$

where i_l is the height of the polarographic wave observed, $S = RT/nF$, and i' is measured with the foot of the wave as the zero line.

In the case where only the oxidized form of the electroactive component is present in the bulk of the solution, combination of eqns. (2) and (3) gives:

TABLE 1

List of symbols

C_O^*	Concentration of oxidised form (Ox.) in bulk of solution
C_R^*	Concentration of reduced form (Red.) in bulk of solution
D_O	Diffusion constant of Ox.
D_R	Diffusion constant of Red.
E	Voltage of the dropping mercury electrode
F	Faraday's constant
j, m	Integers
n	Number of electrons transferred between Ox. and Red.
q	Area of mercury drop at time t
R	Gas constant
T	Absolute temperature
t	Time elapsed since start of mercury drop
V	Half period of square-wave potential

TABLE 2

Factors for instantaneous current in Kalousek polarography

Number of pulses per droplife of 960 ms	F_1 measurement at E_1	F_1 measurement at E_2	F_2 measurement at E_1	F_2 measurement at E_2
2	1.29	1.09	0.65	−1.83
3	1.18	1.07	0.98	−2.10
4	1.13	1.05	1.24	−2.33
6	1.09	1.04	1.65	−2.72
12	1.05	1.03	2.57	−3.62
16	1.05	1.03	3.05	−4.09
24	1.04	1.03	3.86	−4.89
32	1.03	1.02	4.54	−5.56
48	1.03	1.02	5.67	−6.70
96	1.02	1.02	8.23	−9.26

$$i_l = A \,.F_2\,.C_O^* \tag{4}$$

and

$$E_{\frac{1}{2}} = (RT/nF) \ln (D_R/D_O)^{\frac{1}{2}} + E_0 \tag{5}$$

EXPERIMENTAL

Chemicals

Cadmium chloride (Analar), potassium chloride (Merck, reagent grade), lithium chloride (Merck, Suprapur) and mercury (Drijfhout, polarographic grade) were used as received. Tetraethylammonium perchlorate (TEAP, Eastman) was recrystallized from ethanol (Merck, reagent grade) and dried in vacuo at 40°C.

Apparatus

The polarographic system consisted of a Radiometer polarographic stand (type E64) equipped with a drop-life timer (type DLT1). A PDP-11/10 (Digital Equipment Corp.) online computer system with 16K memory was used with a dual disk drive (RKO5), LPS system for A/D and D/A conversion, and a Decwriter. The specifications of the LPS system were: 12 bits A/D converter with dual sample and hold option, range ± 5 V and ± 1 V; 12 bits D/A converter, range ± 10 V at 10 mA max. Also available in this LPS system is a programmable real-time clock with rates up to 1 MHz. A Keithley model 301 operational amplifier was used to convert the polarographic current to the A/D converter input voltage.

A schematic diagram of the system is given in Fig. 1. The dropping mercury electrode (DME) had the following characteristics: m = 2.58 mg s^{-1}, t = 3.4 s in 1 M KCl. Throughout all experiments a mercury pool was used as auxiliary electrode as well as reference electrode. For the lithium determinations, the DME characteristics were m = 1.93 mg s^{-1}, t = 4.83 s (1 M KCl, open circuit).

Polarographic procedure

Treatment of the samples and the cell was as described earlier [5]. For concentrations from 10^{-4} to 10^{-5} M, averaging over 8 successive scans was applied before curve-fitting the data. For the cadmium determinations, the background electrolyte was 1 M potassium chloride, whereas potassium and lithium were determined in the presence of 0.1 M tetraethylammonium perchlorate. In all determinations the staircase potential step was 10 mV per drop time. The cadmium determinations were carried out with E_1 at —0.200 V and E_2 ranging from —0.200 to —0.800 V. For the potassium and lithium determinations these values were —0.800 and —0.800 to —2.500; and —2.000 with —2.000 to —2.700 V, respectively.

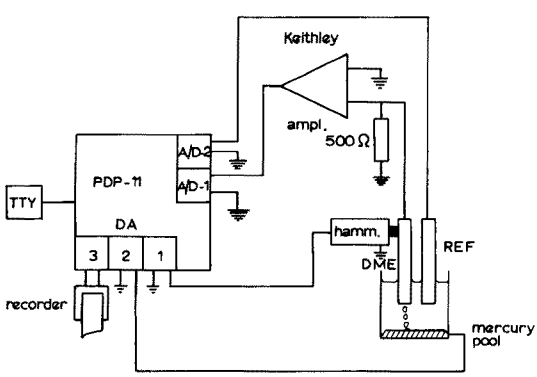


Fig. 1. Schematic diagram of equipment for computer-controlled Kalousek polarography.

COMPUTER PROGRAMS

The software for the fully computerized Kalousek polarography was developed as a set of several separate programs performing the following functions: control of the experiment and measurements; reduction of the experimental data to current versus square-wave amplitude tables; averaging of the current versus square-wave amplitude curves over successive scans; display of the polarograms on a strip-chart recorder or in digital form; curve-fitting of the polarograms to obtain limiting current, half-wave potential, and slope of the log plot.

The programs were to run under a 3-user real-time operating system with background facilities developed in this laboratory for the PDP-11/10. Under this operating system the real-time user has access to a software clock with a period of 120 ms derived from a line frequency clock and a programmable clock with rates up to 1 MHz. The programs for the control of the experiment and the display of the polarograms are written for the real-time environment while the other programs run in the background.

Kalousek polarography is mainly a matter of timing with regard to the pulses applied to the cell and the current measurement. In the real-time program that performs the Kalousek experiment, the square-wave voltage applied to the cell is generated under interrupt with the aid of the programmable clock from the LPS system. The number of pulses is restricted to the following set: 96, 48, 32, 24, 16, 12, 6, 4, 3 and 2 pulses per drop time. Current and cell voltage measurements are done simultaneously, but only during a time slice of 40 ms each 120 ms because of the nature of the operating system. They are initiated by the software clock. Current and cell voltage data are written to disk per drop lifetime. The most important routine in the program is the interrupt routine for the software clock; a flow diagram is given in Fig. 2. The main program senses continuously the flags MFLAG and DISKFLAG which are set by this interrupt routine.

When MFLAG is set, the main program performs the current and cell voltage measurements at a rate of about 2.5 kHz. The setting of DISKFLAG by this clock routine causes the main program to transfer the data from memory to disk. The software clock routine also controls the value of the pulse amplitude (VPULSE). The LPS-clock routine uses this value in combination with the value of the pulse base potential to generate square-wave voltage of the cell.

Timing diagrams for the square-wave voltage applied to the cell and the current measurements for 4 and 96 pulses per second are given in Figs. 3 and 4, respectively. Figure 5 shows the cell voltage versus time for 4 pulses/drop over the first 5 drops.

The next program reduces the current and cell voltage data acquired per drop in one or more 40-ms time windows to 3 values, i.e. current at the center of the last negative-going pulse half, current at the center of the last positive pulse half, and cell voltage at the center of the last pulse. To reduce

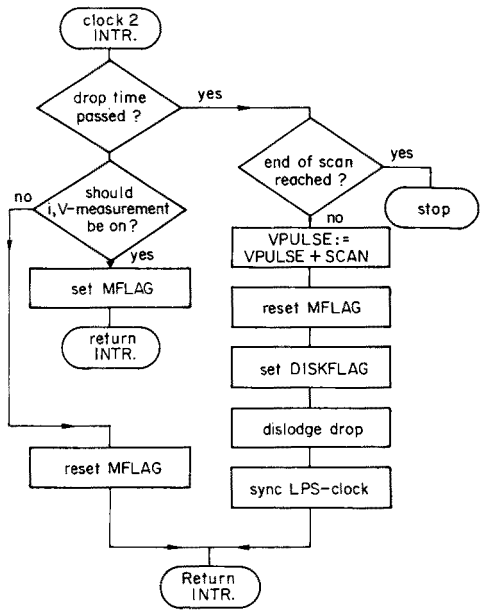


Fig. 2. Flow diagram of program for Kalousek polarographic experiment.

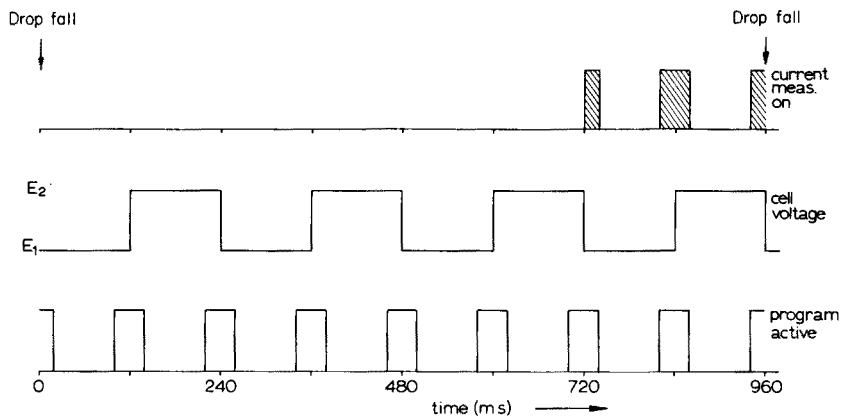


Fig. 3. Timing diagram for cell voltage generation and current—voltage measurements for 4 pulses per drop time of 960 ms.

the influence of line noise on the data, current values are averaged over that part of the half period where the charging current has decayed.

If required, the next program of the set can accumulate the constructed tables comprising the polarograms for averaging over a preset number of successive scans.

When the final polarograms are available, they are displayed by a real-time program on a chart recorder by outputting the current values normalized to

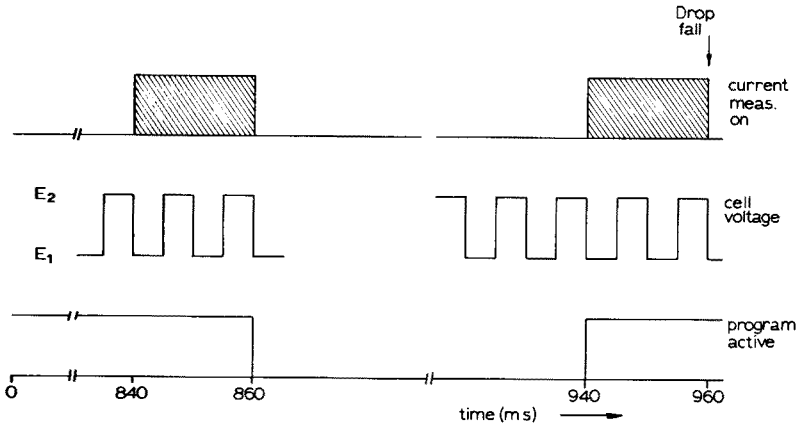


Fig. 4. Timing diagram for cell voltage generation and current—voltage measurements for 96 pulses per drop time of 960 ms.

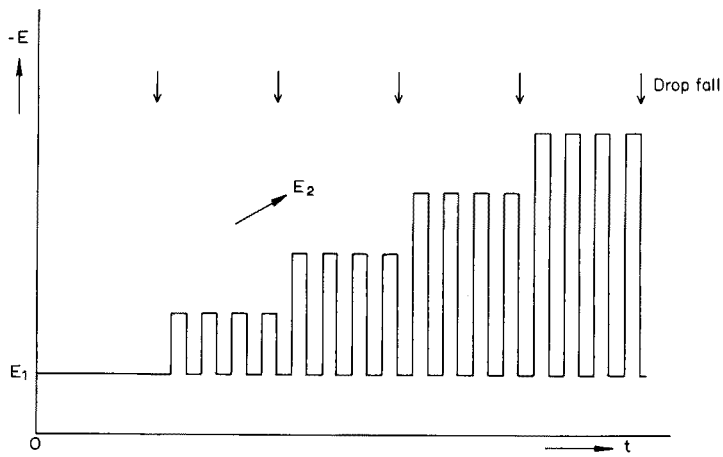


Fig. 5. Cell voltage versus time for 4 pulses/drop over the first 5 drops.

fit the D/A converter connected to the recorder at fixed time increments. The data can also be presented in tabulated numerical form.

Finally the polarographic data are processed by a 3-parameter curve-fitting program to obtain limiting current, half-wave potential, and slope of the log plot. This program is the same as described earlier for the processing of sampled d.c. polarograms [5].

A block diagram of the various stages is given in Fig. 6.

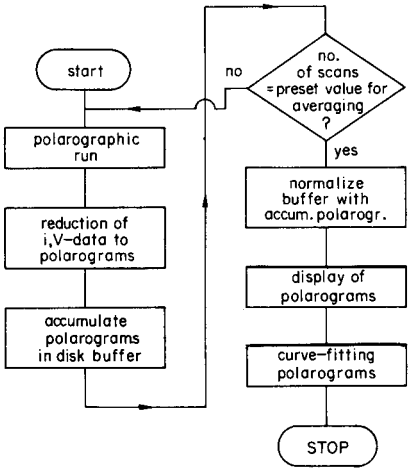


Fig. 6. Flow diagram for complete set of programs for fully computerized Kalousek polarography.

RESULTS AND CONCLUSIONS

All determinations were carried out at a drop lifetime of 960 ms.

Table 3 shows the limiting currents measured at 48 and 16 pulses per drop for various concentrations of cadmium, potassium and lithium; there is a good linear relationship between limiting current and concentration for these Kalousek type II (Heyrovsky's [6] notation) measurements. In all cases the pulse base potential chosen was far more anodic than the halfwave potential, whereas the pulse amplitude was changed in 10 mV steps well into the cathodic limiting range.

For cadmium and potassium, the type I measurements also give useful results (Table 4). For lithium, the type I wave is completely masked by the hydrogen ion discharge wave.

The increase in sensitivity with higher pulse rates is demonstrated in Table 5 for potassium; the mean i_l^{I}/c and i_l^{II}/c values for the various numbers of pulses per drop, divided by the corresponding theoretically calculated F_2 values, are also given. The agreement between theory and experiment is better for the type I measurements than for the type II measurements. Despite the irregularities for 48 and 16 pulses per drop in the type II measurements, the theoretically calculated F_2 values serve to calculate the sensitivity that can be obtained at a given Kalousek frequency. The accuracy of the determinations decreases at pulse rates of 32 pulses per drop and higher; this may be caused by the relatively greater uncertainty in the location of the current measurements at the center of the pulse half periods.

As the reduction of oxygen at the DME is strongly irreversible, there is no interference from oxygen in the recording of type II waves. Table 6

TABLE 3

Kalousek type II limiting currents for Cd^{2+}, K^+ and Li^+

Cadmium					Potassium					Lithium				
	48 pulses/drop		16 pulses/drop			48 pulses/drop		16 pulses/drop			48 pulses/drop		16 pulses/drop	
Conc.[a]	i_l[b]	i_l/c[c]	i_l[b]	i_l/c[c]	Conc.[a]	i_l[b]	i_l/c[c]	i_l[b]	i_l/c[c]	Conc.[a]	i_l[b]	i_l/c[c]	i_l[b]	i_l/c[c]
32.0	7.94	2.48	2.18	0.68	38.4	7.55	1.97	2.11	0.55	38.4	5.38	1.40	1.36	0.35
24.0	6.38	2.65	1.73	0.72	29.1	6.85	2.35	1.71	0.59	29.1	4.17	1.43	1.02	0.35
16.0	4.63	2.89	1.11	0.69	19.6	3.91	1.99	1.10	0.56	19.6	2.78	1.41	0.70	0.36
8.0	2.19	2.73	0.58	0.72	15.0	3.69	2.46			15.0	2.14	1.42	0.50	0.33
4.0	0.89	2.22	0.31	0.78	10.0	2.82	2.82	0.56	0.56	10.0	1.21	1.21	0.35	0.35
2.0	0.52	2.60	0.13	0.65	5.0	1.36	2.72	0.31	0.62					
Mean		2.57		0.71			2.38		0.57			1.37		0.35
σ		8%		6%			14%		5%			7%		3%

[a]($\times 10^{-5}$ M). [b]μA. [c]($\times 10^4$ μA mol^{-1}).

TABLE 4

Kalousek type I limiting currents for Cd^{2+} and K^+

Cadmium					Potassium				
	48 pulses/drop		16 pulses/drop			48 pulses/drop		16 pulses/drop	
Conc.[a]	i_l[b]	i_l/c[c]	i_l[b]	i_l/c[c]	Conc.[a]	i_l[b]	i_l/c[c]	i_l[b]	i_l/c[c]
32.0	11.94	3.73	6.48	2.03	38.4	8.48	2.20	5.53	1.44
24.0	8.62	3.59	4.84	2.01	29.1	8.69	2.98	4.25	1.46
16.0	6.31	3.94	3.27	2.04	19.6	4.96	2.53	2.80	1.42
8.0	3.18	3.97	1.58	1.97	15.0	3.87	2.58		
4.0	1.36	3.40	0.84	2.10	10.0	2.79	2.79	1.38	1.38
2.0	0.69	3.46			5.0	1.69	3.38	0.71	1.42
1.0	0.30	3.00							
Mean		3.58		2.03			2.74		1.42
σ		9%		2%			15%		2%

[a]For footnotes, see Table 3.

TABLE 5

Sensitivity for Kalousek type I and II measurements at various pulse rates

No. of pulses per drop.[a]	Potassium				Cadmium			
	$\frac{i_l^{I}}{c}$[b]	$\frac{i_l^{II}}{c}$[b]	$\frac{i_l^{I}}{cF_2^{I}}$[c]	$\frac{i_l^{II}}{cF_2^{II}}$[c]	$\frac{i_l^{I}}{c}$[b]	$\frac{i_l^{II}}{c}$[b]	$\frac{i_l^{I}}{cF_2^{I}}$[c]	$\frac{i_l^{II}}{cF_2^{II}}$[c]
48	2.74 ± 0.41	2.38 ± 0.33	0.41	0.42	3.48 ± 0.11	2.32 ± 0.19	0.52	0.41
32	2.00 ± 0.05	1.19 ± 0.10	0.36	0.26	2.79 ± 0.17	1.55 ± 0.06	0.50	0.34
24	1.84 ± 0.05	1.10 ± 0.01	0.38	0.29	2.28 ± 0.19	1.25 ± 0.05	0.47	0.32
16	1.42 ± 0.03	0.57 ± 0.03	0.35	0.19	2.05 ± 0.19	0.66 ± 0.02	0.50	0.22
12	1.07 ± 0.05	0.70 ± 0.01	0.30	0.27	1.69 ± 0.10	0.79 ± 0.03	0.47	0.31
6	0.96 ± 0.03	0.47 ± 0.02	0.35	0.28	1.22 ± 0.11	0.52 ± 0.03	0.45	0.32
3	0.75 ± 0.02	0.22 ± 0.02	0.36	0.23				

[a]Drop time, 960 ms. [b]($\times 10^4$ μA mol^{-1}). [c]($\times 10^4$ μA mol^{-1} s$^{\frac{1}{2}}$).

shows the results of cadmium determinations of type II with and without prior removal of oxygen. Within the accuracy limits there is no difference in the limiting current data. However, there is a significant shift in half-wave potential from 651 ± 3 mV for the deoxygenated samples to 684 ± 3 mV in the presence of oxygen. A possible explanation is a reaction between cadmium ions and reaction products of the reduction of oxygen.

The improvement of the signal-to-noise ratio by averaging over successive scans, and the resulting increase in the accuracy of the determination of the limiting current, was found from a series of measurements of 2.91×10^{-4} M K^+ in 0.1 M TEAP. For averaging over 8 scans, i_l^{II} was 3.40 μA ± 3.0%; averaging over 64 scans gave 3.43 μA ± 1.6%.

TABLE 6

Kalousek type II limiting currents for cadmium at 48 pulses per 960 ms drop time with and without prior removal of oxygen

Conc.[a]	Oxygen present		Oxygen absent	
	i_l^{II}[b]	i_l^{II}/c[c]	i_l^{II}[b]	i_l^{II}/c[c]
38.4	9.04	2.35	8.47	2.20
29.1	6.90	2.37	7.04	2.41
19.6	4.43	2.26	4.69	2.39
15.0	3.46	2.30	3.24	2.16
10.0	2.30	2.30	2.43	2.43
5.0	1.15	2.30	1.22	2.44
Mean		2.31		2.33
σ		2%		5%

[a]For footnotes, see Table 3.

Averaging over successive scans can also be used to lower the detection limit of the method. As there is practically no depletion of the bulk of the solution for Kalousek measurements at higher frequencies, time is the only limiting factor in this process. Some examples are given in Table 7 for cadmium determinations in the 10^{-6} M range for 48 pulses/drop. It will be clear that the method becomes impracticable for concentrations below 3×10^{-6} M.

As to the duration of the complete cycle of data acquisition and data processing, the following figures are typical values: experiment, 1—2 min for one scan; recording of the polarogram, 0.5—1 min; and the curve fitting, 1—3 min for one wave. This amounts to 1—6 min for the complete cycle.

TABLE 7

Determination of cadmium by Kalousek type II polarography with averaging over successive scans at 48 pulses per 960 ms drop time

Cd added ($\times 10^{-6}$ M)	Cd found ($\times 10^{-6}$ M)	Deviation (%)	No. of scans averaged
6.89	6.42	−7	47
3.33	3.53	+6	64
1.48	1.05	−29	128
1.00	1.77	+77	237

The author thanks Ms. B. Verbeeten-van Hettema for preparing the manuscript, Mr. R. H. Arends for making the drawings, and Prof. Dr. Ir. E. A. M. F. Dahmen for his interest and encouragement.

REFERENCES

1 I. Růzić, J. Electroanal. Chem., 39 (1972) 111.
2 W. F. Kinard, R. H. Philp and R. C. Propst, Anal. Chem., 39 (1967) 1556.
3 T. Kambara, Bull. Chem. Soc. Jpn., 27 (1954) 523, 527, 529.
4 J. Koutecký, Collect. Czech. Chem. Commun., 21 (1956) 433.
5 M. Bos, Anal. Chim. Acta, 81 (1976) 21.
6 J. Heyrovský and J. Kůta, Grundlagen der Polarographie, Akademie Verlag, Berlin, 1965, p. 454.

Analytica Chimica Acta, 103 (1978) 379—388
Computer Techniques and Optimization

A MICRO-COMPUTER SYSTEM FOR POTENTIOMETRIC STRIPPING ANALYSIS

T. ANFÄLT‡ and M. STRANDBERG*

Department of Analytical Chemistry, University of Gothenburg, Fack, S-402 20 Göteborg (Sweden)

(Received 21st April 1978)

SUMMARY

In potentiometric stripping analysis, trace metals are first amalgamated into a mercury film electrode and then oxidized back into solution by an oxidant. The potential vs. time curve is recorded. Microcomputer-controlled equipment for automation of the potentiometric stripping analysis is described. The higher resolution of time compared to manual equipment increases the sensitivity of the method. A resident, 3-kB BASIC interpreter with routines for handling the hardware makes the programming simple and the instrumentation flexible for changes in the analytical scheme. Printer and keyboard are incorporated in the equipment which makes the whole analyser compact.

Trace metal analysis has become increasingly important in recent years. A new electrometric approach, potentiometric stripping analysis, developed by Jagner et al. [1, 2], simplifies the instrumentation considerably. In this technique metals are first electrolysed into a mercury thin-film electrode. The electrode is then disconnected from the voltage source and the reduced metals in the film are re-oxidized by an oxidant into the solution. The resulting potential curve (Fig. 1) is measured. The time for each metal to be stripped from the electrode is proportional to the initial concentration in the sample and electrolysis time. The elements available for potentiometric stripping analysis are the same as for anodic stripping voltammetry (Table 1). The primary advantage over anodic stripping analysis is the simplicity in the equipment, which is due to the fact that time instead of current is measured. The basic set-up for potentiometric stripping analysis is shown in Fig. 2. It consists of a high impedance amplifier connected to the working and the reference electrodes and an independent voltage source connected to the counter and working electrodes. The voltage source can be disconnected from the working electrode by a switch.

The principal drawbacks with this simple equipment are the difficulty of controlling the plating potential and time, and the low resolution of time

‡Present address: BIFOK, Box 7004, S-191 07 Sollentuna (Sweden).
*Present address: Astra Pharmaceuticals AB, S-151 85 Södertälje (Sweden).

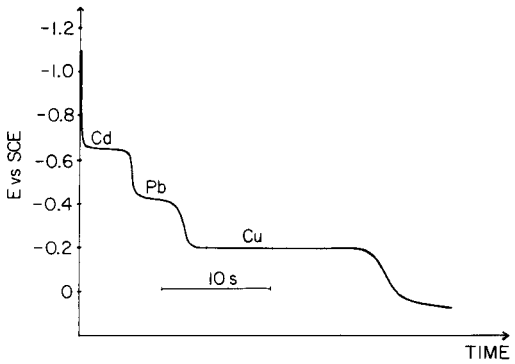


Fig. 1. Potential curve obtained with a plating time of 5 min and a plating potential of −1200 mV. The concentrations are 2 ppm Cd, 2 ppm Pb and 0.2 ppm Cu and the sample is not deaerated.

TABLE 1

Elements suitable for potentiometric scanning analysis

Element	Optimum pH range	Optimum plating potential, E (V vs. SCE)	Scanning potential $E_{1/2}$ (V vs. SCE)
Bi(III)	0—0.5	−0.8— −1.0	−0.15
Cu(II)	0—6	−0.7— −0.95	−0.25
Pb(II)	0—6	−0.8— −1.4	−0.50
Tl(III), Tl(I)	0—6	−0.9— −1.4	−0.65
Cd(II)	0—6	−1.0— −1.4	−0.65
Zn(II)	2—6	−1.3— −1.4	−1.05

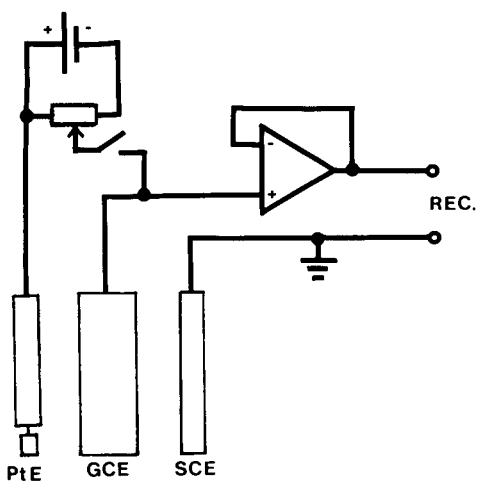


Fig. 2. Basic set up for p.s.a. with a glassy carbon electrode (GCE).

from an x—t recorder. Both these drawbacks can be simply overcome by using a computer to control the analysis. Furthermore, because the computer can also deal with things like sample changing, standard additions and evaluation of the result, the total analysis time can be considerably reduced. The evolution of micro-computers has made it possible to build instruments of high complexity at a relatively low cost. This paper describes a microcomputer-controlled potentiometric stripping instrument and the advantages obtained by building the instrument around a microcomputer.

CONSTRUCTION OF THE INSTRUMENT

Digital section

It is very time-consuming to build a computer from the "chip" level; a single-board computer, the INTEL SBC 80/10 (Intel Corp.) [3] was used. The SBC 80 system is very flexible; it is possible to choose among several types of processor cards and expansion cards for memory, digital input/output and analog input/output. The SBC 80/10 card contains, apart from an 8080A processor, 1 kB of read/write memory, sockets for 4 kB of programmable read-only memory (PROM), serial input/output, 48 bits of programmable input/output lines and 1 level of interrupt.

A card to give the additional functions needed for potentiometric stripping analysis (p.s.a.) was designed to fit into the SBC 80 multibus system. This card contains a bus interface, 2 kB of read/write memory, sockets for 4 kB of PROM, 48 lines of programmable input/output lines, 8 relays, and a 12-bit digital-to-analog converter (DAC) (AD562). It is opto-coupled to the computer to isolate it galvanically from the operational amplifiers so as to prevent current passing between the counter and reference electrodes. The use of all input/output lines from both cards is shown in Table 2.

A line-frequency clock, constructed from a phase-locked loop and a monostable multivibrator, was connected to the interrupt request line to give a 20-ms interval timer. The two 12-bit analog-to-digital converters (ADC) are of the integrating type with a conversion time of 20 ms (Teledyne 8702). They are triggered automatically by the interrupt request line. The resolution is 1 mV/bit for ADC 1 and 4 mV/bit for ADC 2, giving ranges of ± 2.048 V and ± 8.192 V, respectively. The DAC has the same resolution and range as ADC 2.

TABLE 2

Input/output lines of the SBC 80/10 and p.s.a. cards

Card	Port	Use	Card	Port	Use
SBC 80/10	1	Printer output and control	P.s.a.	7	DAC
	2	—		8	Keyboard
	3	Display output		9	Least significant bit of DAC
	4	ADC no. 2		10	Keyboard and printer "flags"
	5	ADC no. 1		11	Relays
	6	Least significant bits of ADC nos. 1, 2		12	—

A keyboard (GRI 753) and printer (Olivetti nip 18) were incorporated into the system, to achieve independence from an external terminal. An external terminal can, however, be connected to the serial input/output which also can be used for communication to other computers. A block diagram of the equipment is shown in Fig. 3.

Analog section

The analog parts of the equipment are shown in Fig. 4. The metals in the mercury film can be oxidized either by an oxidant in the solution or by a current passing through the electrodes; it is therefore essential to make the current drawn through the amplifier negligible compared to the oxidation in the solution. Therefore a high-impedance FET amplifier (LF 356) was chosen with an input impedance of 100 Gohm. The differentiating amplifier was added to the system to simplify the evaluation of time between each potential step for the elements. For the convenience of the operator, a panel meter connected to the high-impedance amplifier was added to the instrument to monitor the potential between the working and reference electrodes. With this arrangement it is possible to discover errors such as electrodes which are not properly connected or wrong polarity on the DAC, etc.

Software

Potentiometric stripping analysis is a new technique for which the optimal analytical conditions have still not been fully explored. For optimization,

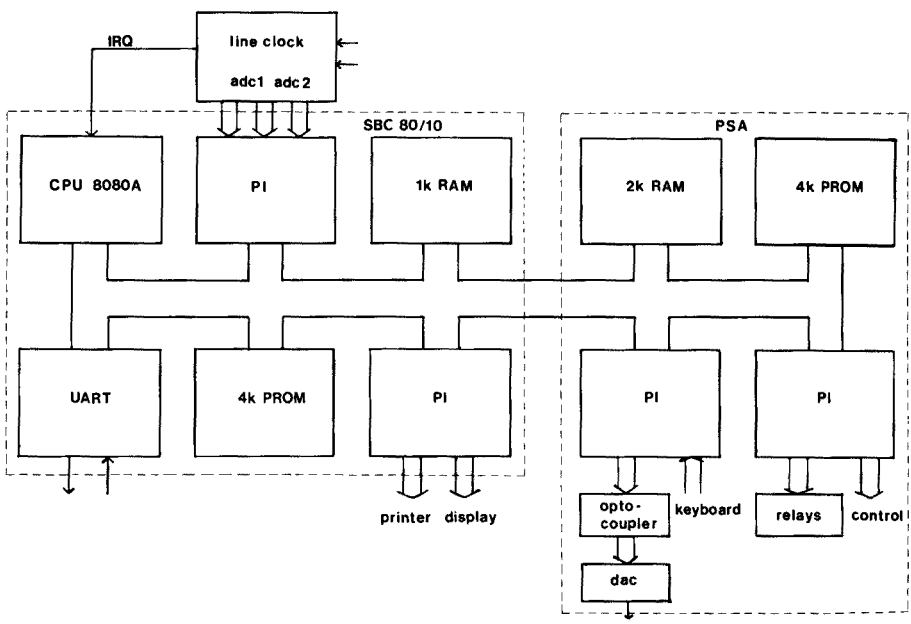


Fig. 3. Block scheme of the instrument. PI, parallel interface; IRQ, interrupt request. Only 4 K Bytes of the PROM capacity is used. Bustranceivers are excluded from the scheme.

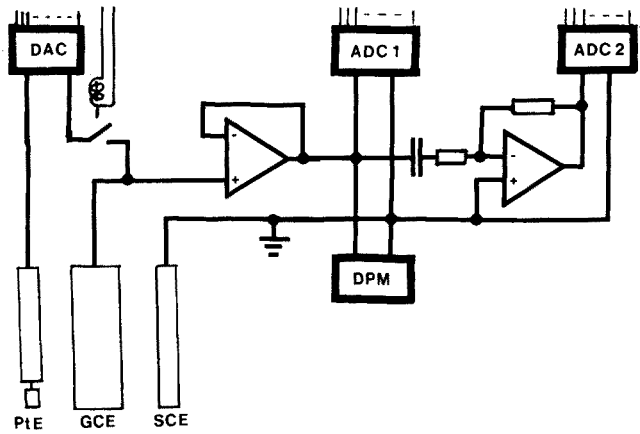


Fig. 4. The analog section of the instrument. The operational amplifiers are NS LF 356. The digital panel meter and the differentiating amplifier are optional.

maximum flexibility in the instrument is needed. Flexibility in changing the analysis scheme calls for an interpretative computer language such as BASIC to give the flexibility. However, one of the drawbacks of BASIC is the relatively low execution speed. Partly to overcome this drawback, and because no complex mathematics is involved in the evaluation of data, a small integer-based BASIC was chosen from the INTEL user library; this had been written by Li Chen Wang and modified to 8080-code by Rauskolb. Integer word length is 16 bits, i.e. numbers between −32767 and +32768 can be used. Bench-mark tests were done to compare it in speed to other 8-bit machine BASICs. Some of the statements, e.g. a FOR · · · NEXT loop, showed 2—3-fold increase in speed compared to conventional BASICs, whereas others like IF showed no great difference in execution time. The elements of this BASIC are essentially the same as in other BASICs. New routines were implemented to utilize the hardware in BASIC programmable form.

The memory map in Table 3 shows that the interpreter including instrument handlers occupies less than 3 kB of read-only memory. The BASIC program code stored in the active read/write memory can use 3 kB shared with data collected and stored by means of the indexed variable @( ). In most applications, it will give a data storage area of about 2 kB, i.e. about 1000 data can be stored in the BASIC program.

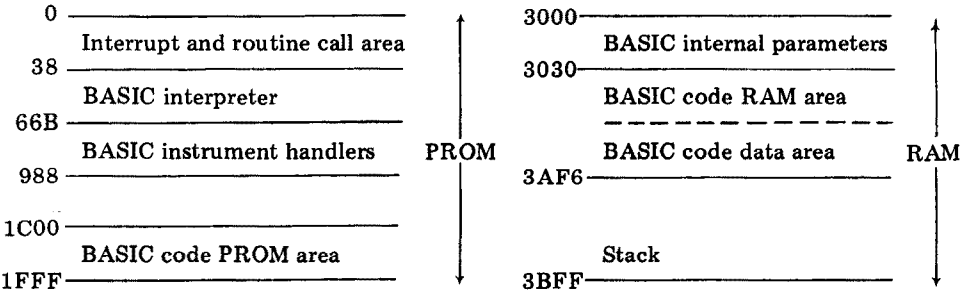
A new command, POP, was added to the BASIC, which moves a pre-programmed BASIC program from the read-only memory to the BASIC working area. This feature facilitates the starting-up procedure if the equipment has been disconnected from the mains voltage, because the read/write memory is "volatile".

The programs are loaded into the read-only memory by using another new command, PUSH, which can only be used under the control of the INTEL ICE-80 in-circuit emulator.

TABLE 3

The organization of the memory. The limit between BASIC code area and BASIC data area is floating and depends on the size of the BASIC program

Hexadecimal address



Functions added to BASIC

All functions added to the BASIC program are summarized in Table 4. The function ADC (i) converts the mV value from the *i*th analog-to-digital converter to the BASIC. The value has been read previously by the interrupt routine, which makes an interrupt each 20th millisecond (see Fig. 5). The ADC (i) function is limited, to ensure that it is impossible to make double readings of the same converted mV value, by means of a software flag. The interrupt routine also increments the 20-ms interval timer which can be set or reset by the TIC (i) function. Since the length of the internal integer is set at 16 bits in this BASIC, the clock span is only about 10 min. If longer times are needed, the clock has to be reset. The DAC (i) routine outputs *i* mV to the digital-to-analog converter. The DAC can either output the voltage to the electrodes or to an external recorder, which is controlled from the RNR (i)

TABLE 4

Additional functions in BASIC

Function	Parameter	Action
ADC (i)	$i = 1$ or 2	Reads mV value from ADC 1 or 2
TIC (i)	$i = 1$ or 2	Zeroes or reads the 20-ms interval times
DAC (i)	$-8191 < i < 8188$	Output i mV to DAC
RNR (ij)	$i = 0$ or 1 $1 < j < 8$	Sets or resets relays
TER (i)	—	Initiates the serial input/output interface
SET	—	Executes functions which follow after set
POP	—	Transfers a program from read-only memory to the active read/write memory
PUSH	—	Transfers a program under ICE-80 control from the active read/write memory to memory which will be read only after simulation.

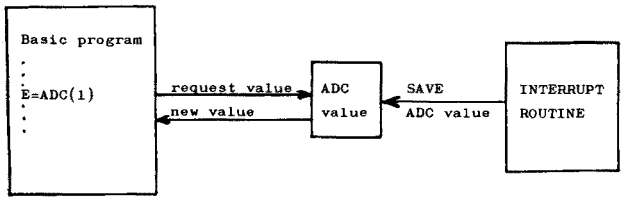


Fig. 5. The interdependence of the interrupt routine and the BASIC function ADC (i).

routine. This routine also sets or resets other relays or digital signals on request from the BASIC program, to enable additional equipment, e.g. automatic burets and samplers to be used in the system. The TER (i) function enables the serial interface to be re-initialized to different baud-rates to suit any external terminal. All functions echo back to BASIC a value from the routine, which often is unnecessary. This can be prevented by using the added statement SET followed by the function.

Procedure

The glassy carbon electrode must be pre-plated with a 250-ppm mercury(II) nitrate solution acidified with nitric acid to pH 3 for 2 min at −0.5 V vs. SCE. This is done once a day or less depending on how much the electrode is used. Positive applied voltages must not be used because the electrode then degenerates drastically. Only polishing with, e.g., diamond paste and replating of the mercury film will then restore the electrode. The BASIC program for controlling the analytical scheme is divided into the same subsections as the whole analysis. The first section which inputs parameters such as the desired electrolysis time and potential is followed by putting the potentiostat into use.

The potentiostat can be programmed as shown in Table 5. The desired potential is output on the DAC, after which the clock is reset and the DAC is connected to the counter and working electrodes by the RNR function. After four TIC's, the ADC is read and the output value of the DAC is corrected for the discrepancy between output value and read value. The correction term is divided by two to avoid overshooting. The result of the regulation is shown in Fig. 6, where it is also possible to see the typical potentiometric stripping analysis curve at the right edge of the curve.

When the time of electrolysis has expired, the relay connecting the DAC to the electrodes is reset. The clock is immediately set to zero, and values from ADC 1 are stored while ADC 2 is observed for the appearance of a maximum. If this is reached, the time for this maximum to appear is subtracted from the earlier "maximum" time and stored. Also the potential from the stored mV values between the two maxima are calculated and stored to facilitate identification of the element. The procedure is repeated until a selected mV value is exceeded, which is normally +100 mV where no further elements are expected. The result is then printed as the time, or the

TABLE 5

Sample program for p.s.a. (see text)

```
1000 REM * * POTENTIOSTAT * *
1010 REM: IN E, T
1020 REM: OUT NONE
1030 REM: USED S, U
1040 SET DAC (E)
1050 SET TIC (1)
1060 SET RNR (1)
1070 S = E
1080 U = TIC (2)
1090 IF TIC (2) —U < 4 GOTO 1090
1100 S = S + (E—ADC (1))/2
1110 SET DAC(S)
1120 U = TIC (2)
1130 IF U < 50 ☆ T GOTO 1090
1140 SET RNR (11)
1150 RETURN
```

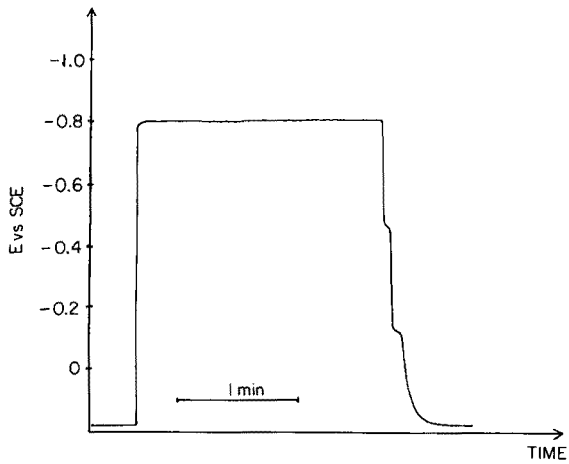


Fig. 6. The resulting potential curve of the program shown in Table 5 with E = —800 mV and T = 120 s in a solution of 2 ppm Pb and 0.2 ppm Cu which was not deaerated.

time is multiplied by a concentration factor, and the "half-wave" potential of the element. Table 6 shows the results obtained for a sample containing 0.2 ppm of cadmium, lead and copper, printed as time intervals.

The instrument has so far been used only for non-deaerated samples where the oxygen oxidizes the metals back into solution. If the samples are first deaerated and then 1—10 ppm mercury(II) chloride is added to the sample, the metals are more slowly oxidized back into the solution, and this gives more than ten times higher resolution for all metals except copper. This decreases the detection limit considerably.

TABLE 6

P.s.a. results for a 0.2-ppm solution of Cd, Pb and Cu

Potential	Time
−188	62
−440	7
−721	12

INITIALS
END POTENTIAL: 0
ELECTROLYSIS TIME: 180
ELECTROLYSIS POTENTIAL: −1000

DISCUSSION

Microcomputers are valuable for controlling analytical instruments and evaluating the results. This can now be done without extreme cost, and changes the traditional analytical chemical procedures. In potentiometric stripping analysis, many advantages can be achieved by integrating the computer into the instrument. The better resolution of time increases the sensitivity of the system, which can be used either to shorten the electrolysis time or to increase the precision and sensitivity of the method. The results can be obtained in many ways, for instance, by the standard addition method where the computer controls a buret which makes the additions, or by evaluating the results from a calibration curve obtained previously. Test evaluations can also be done to estimate concentration levels which then can be used automatically to generate a proper electrolysis time for a chosen precision.

An electrolysis for a fraction of a second followed by data collection can be used to correct the p.s.a. curve for the capacitive unloading of the electrode pair. This is done by subtracting the times of the two curves for each equipotential value from the ADC. It is particularly useful when the concentrations and/or the electrolysis times are small and thus the influence of the capacitive unloading is great.

Full utilization of the possibilities of the computer in implementing complex analytical procedures requires that the programming be done in a high-level language like BASIC. The user cannot benefit from the possibilities of reprogramming the computer if this is not easily accomplished. Many of the commerical instruments equipped with microcomputers are very rigid in this respect, and moreover the software is often very badly documented. The effect is that the user has completely lost control over such vital parts as, e.g., the algorithm for evaluation of the results and has often no possibility whatsoever of changing it to suit his special needs.

Various small PROM-resident BASICs are now available which can quite easily be modified to suit real-time applications of moderate speed. To use BASIC in microprocessor- controlled instruments greatly enhances the flexibility and ease of tailoring the system to special needs.

The widespread use of computers has in many cases led to an overestimation of the possibilities of doing calculations on digitally sampled data. It is important to bear in mind that the digital data represent only the analog reality at single points and often with very limited resolution. This makes digital data susceptible to noise and can make a task which is quite simple in theory, e.g. detecting the inflexion point on a curve, into a very tedious and time-consuming procedure. Where the error introduced permits its use, analog data processing, like squaring, extracting the root mean square, log, antilog, differentiating, etc., is greatly to be preferred.

The combination of analog data processing, which also has the great advantage of being done in "real-time", and digital data processing is a very powerful tool in instrument development.

REFERENCES

1 D. Jagner and A. Graneli, Anal. Chim. Acta, 83 (1976) 19.
2 D. Jagner and K. Åren, Anal. Chim. Acta, 100 (1978) 375.
3 INTEL, SBC 80/10 Hardware reference manual, 98—230B, and Assembly language programming manual, 98—004C, 1976.

Analytica Chimica Acta, 103 (1978) 389—400
Computer Techniques and Optimization

MICROPROCESSOR-ASSISTED HIGH-PRECISION VISCOMETRY

O. S. BORGEN* and O. SANDBU

Trondheim University, The Norwegian Institute of Technology, Physical Chemistry Division, N-7034 Trondheim-NTH (Norway)

(Received 5th May 1978)

SUMMARY

A microprocessor operated system for time measurements for use with a high-precision high-temperature torsion pendulum viscometer, whose oscillations are tracked by electro-optical means, is described. Photodetector signal level transition times, measured to within ± 50 ns, are stored in the microcomputer, and after some calculation and checking of internal consistency, transferred to a PDP-11 minicomputer where the main viscosity computations take place. The system thus represents a simple multiprocessor laboratory network.

A computer-assisted, high-precision, high-temperature torsion pendulum viscometer has been under continuous development at this university [1—4]. The construction of the main viscometer and most of the development of the data reduction program has been done in the Division of Inorganic Chemistry, while the Physical Chemistry Division has been concerned with the on-line computer operation of the viscometer.

The principle of the torsion pendulum viscometer (Fig. 1) is well known and simple. The pendulum, suspended from a torsion wire, either is immersed in, or contains, the liquid whose viscosity is to be measured. After initiation of the torsional motion by a combined torque initiator and brake, the pendulum makes damped oscillations which are tracked by a collimated laser beam reflected from a mirror attached to the torsion wire. A simplified diagram of the optical system is shown in Fig. 2. Two photodiodes in approximately symmetrical locations relative to the rest position of the light beam, give pulses of the order of 1-ms duration when illuminated by the light beam during the oscillation of the pendulum. By precise measurement of the detector crossing times over a number of oscillations, the damping of the pendulum and thus the viscosity of the liquid can be calculated [1, 4].

A typical damped oscillation of the pendulum, as tracked by the deflection of the laser beam from the attached mirror, is shown in Fig. 3(a). When the reflected light beam crosses the detector apertures, pulses for control at the time measurement system are generated: A1 on the leading edge of detector A signals, A2 on the lagging edge of detector A signals, B1 on the leading edge of detector B signals, and B2 on the lagging edge of detector

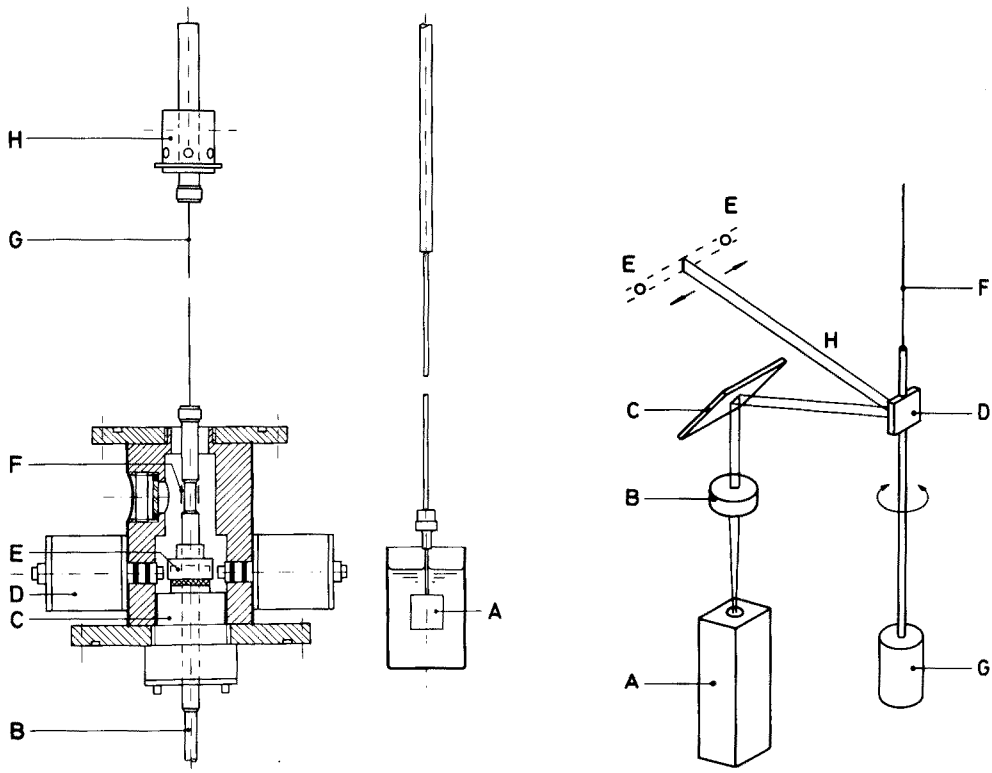


Fig. 1. Viscometer details (vacuum and temperature control systems not shown). (A) oscillating Pt—10% Ir cylinder; (B) torsion pendulum; (C) toroid support; (D) torque initiator magnets; (E) cup-shaped aluminium rotor with internal stationary iron toroid; (F) platinized pendulum mirror; (G) torsion wire; (H) suspension head.

Fig. 2. Simplified optical system. (A) helium—neon laser; (B) optical focusing system; (C) plane mirror; (D) mirror on torsion pendulum; (E) photodetectors (channels A and B); (F) torsion wire; (G) torsion pendulum.

B signals (Fig. 3b). Each of these pulses initiates a time measurement which may be followed by a hardware interrupt to the attached computer. The associated service subroutine transfers the time to storage and, by keeping account of flag bits indicating which of the four signals has caused the interrupt, follows the sequence of detector crossings. In a previously reported version of the viscometer [1], a different time measurement scheme was applied, and a PDP-11/10 computer was used with a KW11-P mainframe clock which was interrogated on detector generated interrupts.

At the successful completion of a run consisting of 20—30 oscillation cycles each of 1—5-s duration, the detector crossing times were available in the PDP-11/10 memory for the data reduction program. This program, written in FOCAL, gave a final report (Fig. 4) after a total run time of some 6 min. For good runs, absolute viscosities with relative standard deviations of about

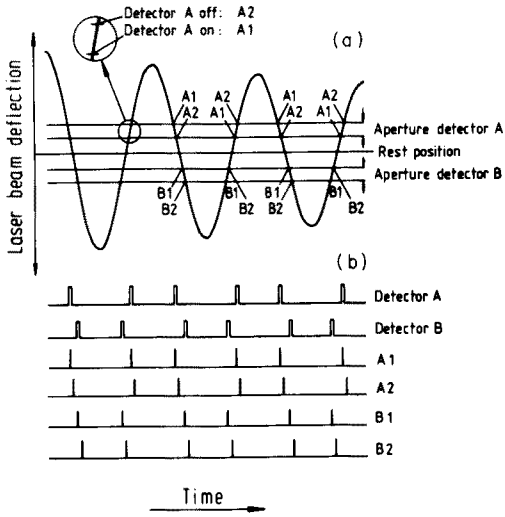


Fig. 3. (a) The deflection of the laser beam is shown over several cycles of the damped pendulum oscillation together with the crossing of detectors A and B; the detector apertures are greatly exaggerated, as is the interdetector distance compared with the deflection amplitude. (b) Time sequence diagram for the detector signals and timer control/interrupt signals A1—B2.

SYSTEM	1	RUN NUMBER	20
DATE	260277	PERIODS INCL.	5 - 28
TEMPERATURE, C	888.80	1000/K	0.8606
RADIUS CM	0.95440	HEIGHT CM	1.91562
ROD RADIUS CM	0.075	ROD IMMERSION CM	1.5
PER. IN GAS MS	1693.08	DAMP. IN GAS	0.001135
INERTIA GCM2	85.530	LIQ. DENSITY	1.508480

		MEAN	ST.DEV	LEFT	RIGHT
PERIOD T (OSC. PEAK)	MS	1698.460	0.003		
PERIOD (OSC. ZERO)	MS	1698.450	0.005		
OSCILLATION ZERO	MM			-0.4472	0.4467
MEAN AMPLITUDES	MM			199.232	201.543
AMPLITUDE DECREASE	MM			106.790	107.941
DAMPING CONSTANT D*10000		230.39	0.04	230.40	230.38
STANDARD DEVIATION *10000				0.05	0.06
CYL. GAS DAMPING D*10000		1.89			
ROD DAMPING D*10000		0.14			
TOTAL DAMP. CORR. D*10000		9.60			
VISCOSITY FROM PERIOD	CP	0.8463	0.0010		
VISCOSITY FROM DAMPING	CP	0.8668	0.0003		
LOG CP FROM DAMPING		-0.0621			

Fig. 4. Printout of results from a representative viscometer run.

0.2% were obtained [4]. A series of measurements on NaCl in the temperature range 820—940°C could be fitted to the model equation: $\eta = A \exp [(B/R)(1/T - 1/TM)]$, through 78 experimental points with a standard deviation of 0.05% [4].

This earlier system has been in successful operation for about four years.

After a number of improvements in the mechanical parts of the viscometer, the stage was finally reached where the standard deviations of the calculated oscillation periods had decreased to the order of 5 μs. As the resolution of the KW11-P mainframe clock is limited to 10 μs, it was decided to design a new computer interface with a resolution of 100 ns, which was judged sufficient by a comfortable margin for any future mechanical and optical improvements in the total system. With the enhanced accuracy, a clock read indirectly through service subroutines could not be used. Major changes were thus necessary. Rather than try to adapt the existing PDP-11 interface, it was decided to design a completely new data collection system based on a microprocessor with appropriate interface. One reason was that the PDP-11 was becoming tied up in other instrumentation projects. Even though most of the computations and reporting of results could still be dealt with by the PDP-11, it seemed better to delegate the main data collection and checking load to external circuitry.

The main impact of microprocessors on laboratory instrumentation — besides the present and potential availability of distributed logical and computing resources — is the ease and very low cost of implementing interface circuitry. In fact, this may be the sole viable distinction between micro and other processors. Development of complex program systems is still very time-consuming if one has no access to one of the more advanced — and expensive — software development systems. For development of the relatively simple program routines needed for distributed microprocessor applications, a cross-assembler or preferably a cross-compiler resident on a larger host computer will often be satisfactory. At the moment, the laboratory data system in this Institute is based on a hierarchal computer system. A PDP-11 computer — with future access to the university data net — will act as a host computer for the main operating and development software and the hardware resources common to the whole laboratory. Communication with LSI-11 and M6800 family satellites will take place over asynchronous channels, while any complex instrument may contain several dedicated microprocessors, each with limited functions, interconnected and tied to the satellites through IEEE-488 buses or asynchronous current loops. The use of such distributed control computers will become even more important when the new generation of microprocessors, e.g. the M6801 containing processor, memory, timing and input-output circuits, becomes easily available.

The system described here represents a very simple application of a mini-microcomputer system.

MICROPROCESSOR SYSTEM

The arrangement adopted consists of a 6800 microcomputer-controlled system (Motorola/American Microcircuits Inc.) which stores the detector crossing times as measured by the separate timer, performs some calculations,

and transmits the results to the PDP-11 on the TTY asynchronous line for final data processing.

The running-in of the system is done with a prototyping board EVK 300 (American Microsystems, Inc.; AMI) which contains an S6800 microprocessor, 4K byte read-only memory (ROM) containing an operating system, 1 K byte random access memory (RAM), 2 K byte erasable read-only memory (EPROM), 3 MC6820 peripheral interface adapters (PIA), 1 MC6850 asynchronous communications interface adapter (ACIA); ancillaries are an internal timer, facilities for loading S6834 512-byte EPROMs, buffered address, control and data lines. Details of the microcomputer large-scale integrated circuits and their programming are given in the manufacturers' literature for the AMI 6800 and M6800 microprocessors.

The MC6800 family of microcomputer circuits was selected because these circuits were in use in other instrumentation projects, and a cross-assembler is available for the MC6800 on a PDP-11/20 computer with magnetic disc storage and an RT-11 operating system.

The cross-assembler [5] is based on a report by Seim [6], who described a set of cross-assemblers generated by macro-statements under the RT-11

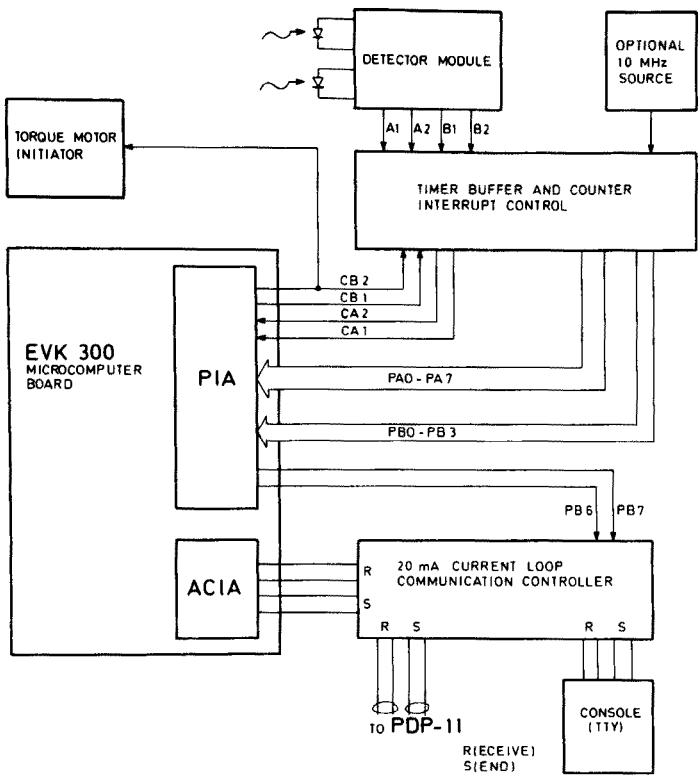


Fig. 5. Block diagram of microprocessor interface.

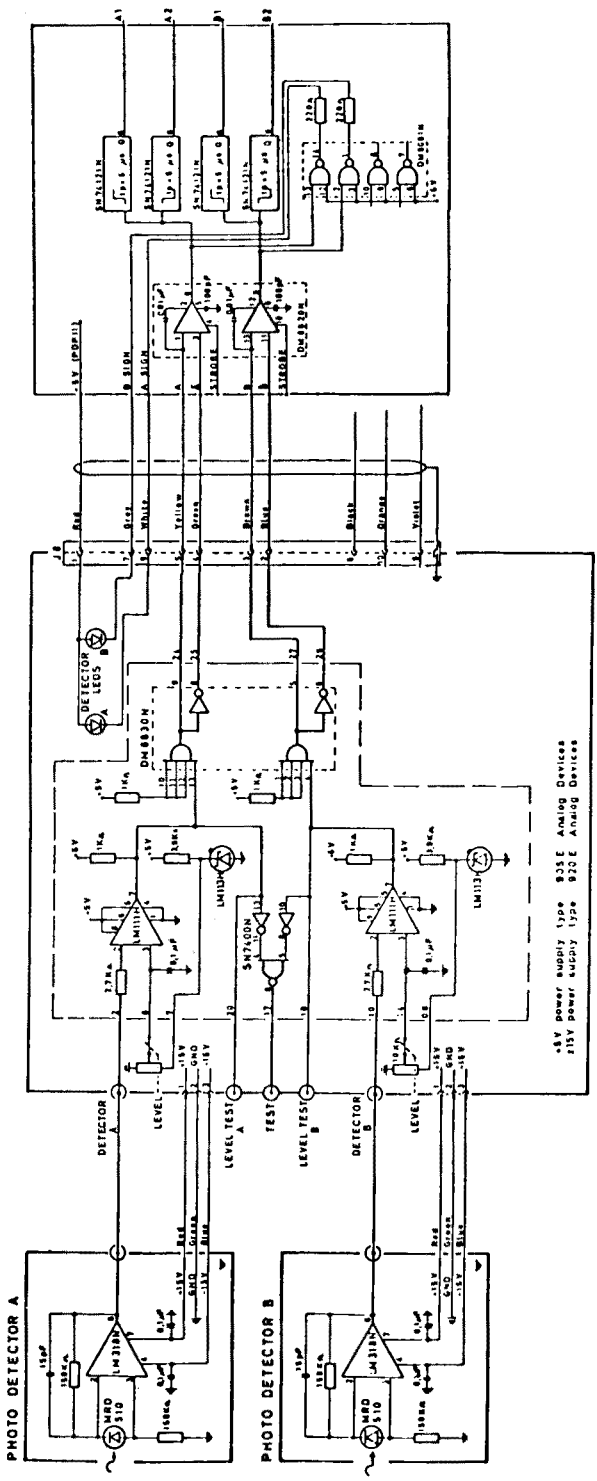


Fig. 6. Photodetector module.

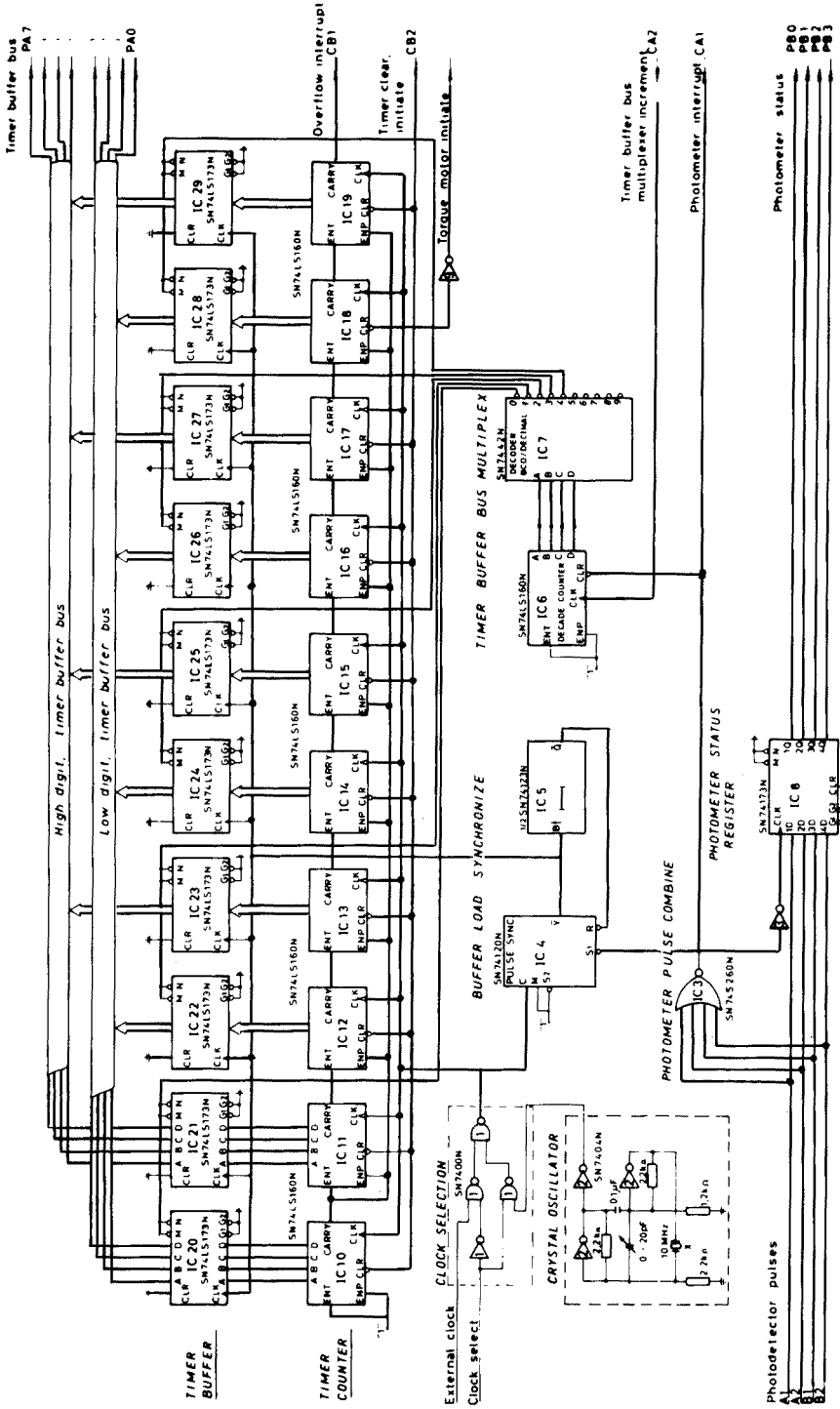


Fig. 7. Timer and interrupt circuits.

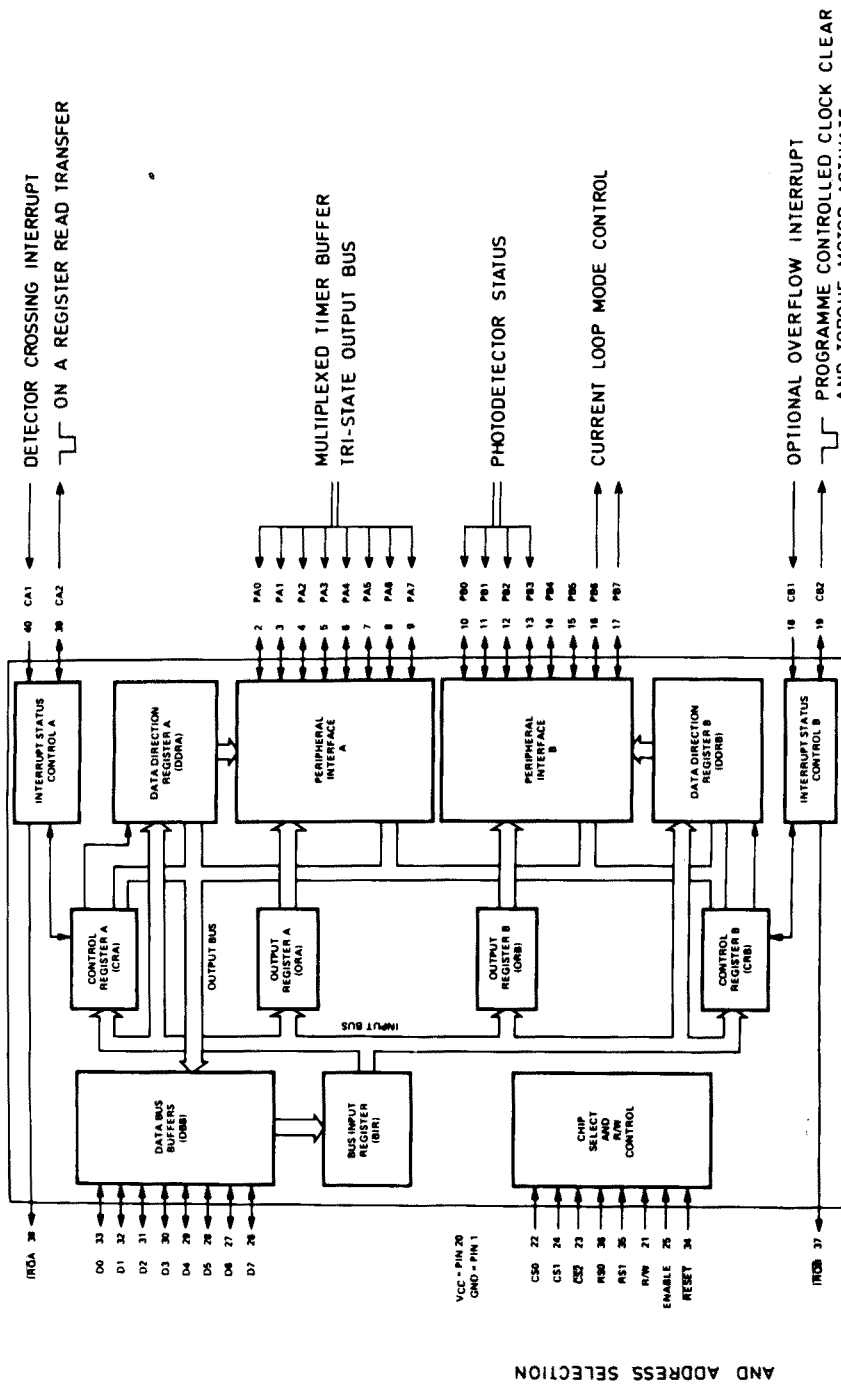


Fig. 8. Connections to the S6820 peripheral interface adapter. (In part reproduced by courtesy of American Microsystems, Inc.)

MACRO assembler. The report described an incomplete cross-assembler containing several errors. The present version of the cross-assembler appears to be fully debugged, and may be obtained from the authors.

A block diagram of the microprocessor-controlled system in its present form is shown in Fig. 5. The main interface modules are described briefly below.

Detector module

The photodetector module (Fig. 6) consists of two channels of photodiode detectors MRD510 with LM318N amplifiers in a temperature-drift compensation circuit and level-determining LM111H comparators. The detector pulses are transferred to the remotely located main interface by the line driver/receiver pairs DM8830N and DM8820N. The latter trigger the SN74121N that generate the final detector output signals A1, A2, B1 and B2. It has recently been found that substitution of fast opto-couplers for the line driver/receiver pairs gives an appreciable reduction in noise level.

Timer and interrupt control

This module (Fig. 7) contains a 10-MHz crystal-controlled oscillator; an external precision 10-MHz source may be substituted for the internal source. The timer counter consists of ten cascaded SN74160 BCD counters (IC 10-19). The counter chain is reset by an initiation pulse on control line CB2. An overflow signal is also available from the CARRY output of the final counter (IC 19) for initiation of interrupts on control line CB1. The timer buffer consists of ten SN74173 D-registers (IC 20-29) with data inputs connected to the output ports of the SN74160 counters (IC 10-19). The TRI-STATE outputs of the SN74173 are bussed, in pairs, to the data lines PA0-PA7 of the PIA.

Any of the photometer pulses A1—B2 combined by a NOR-gate (IC 3) may initiate an interrupt on control line CA1. Simultaneously, the SN74160 multiplex control counter (IC 6) is cleared, and the states of A1—B2 are latched onto the data lines PB0-3 by the SN74173 D-register (IC 8). A SN74120 pulse-synchronizing circuit (IC 4 and 5) activated by the leading edge of the combined photometer signal, will transmit a single inverted pulse from the clock oscillator, loading the timer counter contents into the timer buffer. This transfer will take place between two clock counter increments, defining the time of any photodetector crossing to within one 100-ns interval (with a delay which is essentially constant over any one run).

The timer multiplexer consists of the SN74160 counter (IC 6), which is cleared by any detector pulse as mentioned above, and the SN77442 BCD-decoder (IC 7) which activates the timer buffer register outputs in pairs, connecting them to data lines PA0-PA7. The first read transaction of the PA0-PA7 data line will transfer the stored values of the two first timer buffer registers (IC 20 and IC 21) to the processor. After the transfer, a negative pulse on

CA2 increments the counter, activating the two next buffer registers (IC 22 and IC 23) and so forth, until the complete buffer contents have been transferred.

Peripheral interface adapter (PIA)

The PIA (Fig. 8) is under program control from the microprocessor. The two sets of I/O registers are independently addressed. Any of the lines PA0—PA7 and PB0—PB7 must be programmed for input or output, and are initiated as shown in Fig. 8. The functions of the four control lines (CA1, CB1, CA2 and CB2) must be programmed according to their later functions. In this case, transitions on CA1 and CB1 initiate interrupt sequences. CA1 outputs a negative pulse each time a read transfer on PA0—PA7 takes place. CB2 is under direct program control.

Interface operation (Fig. 5.)

A run is initiated by a programmed negative pulse on control line CB2 which resets the timer counter and activates the torque initiator (which may also be manually operated). During a run, detector crossings cause the generation of the pulses A1—B2 (see Fig. 3). Any of these may cause an interrupt via control line CA1, also transferring the timer counter contents to the timer buffer within the following clock pulse interval, and setting flag bits on the following data input lines: PB0 = 1 for A1, PB1 = 1 for A2, PB2 = 1 for B1, and PB3 = 1 for B2.

The processor reads the buffer contents of 10 BCD digits by 5-byte transfers over the data input lines PA0—PA7. The buffer output is multiplexed onto these lines. For each transfer read, a negative transition on control line CA2 increments the multiplexer control counter, the latter being reset by any of the detectors signals A1—B2.

The 6800 software is quite simple and will not be described. However, by keeping track of the photometer interrupt sequence, the program contains a strong safeguard against viscometer malfunctions. On initiation of the pendulum oscillation in the direction of detector B, any initial B1 and B2 interrupts will be disregarded. The interrupt sequence during a normal run will be A1 A2 A1 A2 B1 B2 B1 B2 A1 A2 . . . . Any deviation from this sequence will cause the run to be terminated. Decaying oscillation amplitudes will be signalled by sequences of the type A1 A2 A1 A2 B1 B2 A1 A2 . . . or A1 A2 A1 A2 A1 A2 . . . , or even an absence of detector interrupts.

After initial calculations depending on the chosen measuring strategy, the 10-digit BCD numbers are transmitted to the PDP-11 as ASCII digits through the console TTY current loop. The use of multi-precision BCD numbers will, for very good reason, limit the amount of preprocessing. The FORTRAN driver program in PDP-11 accepts the data as double precision numbers.

As the timer has a capacity of 999.999 999 9 s, and is reset at the beginnin of any viscometer run, an overflow will rarely occur. However, a timer overflow can be signalled by an interrupt on control line CB1.

The current-loop communications controller permits three modes of interconnection of the PDP-11 and M6800 asynchronous communication ports and the teletype console. The mode selection is under program control of the M6800 through output bits PB6 and PB7. The current-loop switching is done by TTL-compatible reed relays.

By output of selected non-printing characters to the M6800, the PDP-11 may request a change of communication mode. The communication modes (Fig. 9) are as follows:
(a) PB6 = 0, PB7 = 0; normal mode; console connected directly to PDP-11 while M6800 monitors output from PDP-11 to console;
(b) PB6 = 1, PB7 = 0; report mode; full communication between PDP-11 and M6800; console switched to local operation only;
(c) PB6 = 0, PB7 = 1; report mode with logging on console (this is mainly intended as an error-checking facility); all output from M6800 to PDP-11 is available for immediate inspection on console.

CONCLUSIONS

The microprocessor interface has performed satisfactorily during the testing stage. Assessment of the eventual precision of the total system must await an extended period of use, together with attempts to improve the mechanical and, in particular, the optical components. Some uncertainty as to whether

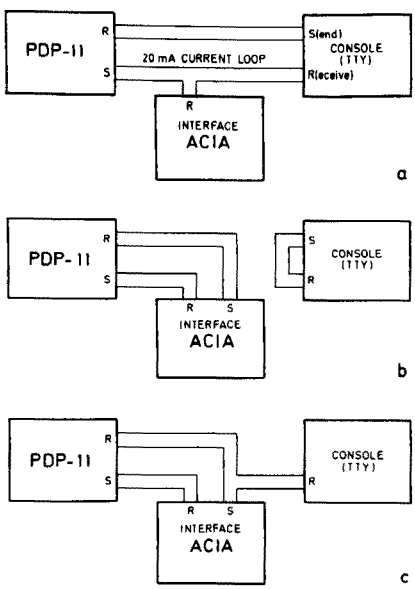


Fig. 9. Simplified block diagram of the current loop communications controller. (a) normal console TTY operation; (b) Interface report mode; (c) Interface report mode with data monitoring on console TTY.

one should measure the time intervals between detector crossings rather than the absolute crossing times may result in later modifications.

As the data transfer takes place at TTY-speed, roughly one 10-digit number will be transferred each second, while the rate of oscillation of the viscometer is one period per 1—5 s. In a later version, therefore, it may be found better to introduce a second asynchronous channel on the PDP-11 dedicated to communication with the microcomputer at a higher rate of transfer.

REFERENCES

1 T. Ohta, O. Borgen, W. Brockner, D. Fremstad, K. Grjotheim, K. Tørklep and H. Øye, Ber. Bunsenges. Phys. Chem., 79 (1975) 335.
2 O. Borgen, D. Fremstad and H. Petersen, Trondheim University, The Norwegian Institute of Technology, Physical Chemistry Division, NIT-PCD Technical Report 53.
3 O. Borgen and H. Petersen, NIT-PCD Technical Report 69.
4 K. Tørklep and H. A. Øye, Ber. Bunsenges. Phys. Chem., submitted.
5 O. Borgen, G. A. Hansen and T. Sparre Olsen, NIT-PCD Technical Report 86.
6 T. A. Seim, Pacific Northwest Laboratory, BNWL-SA-5701 (1976).

Analytica Chimica Acta, 103 (1978) 401—408
Computer Techniques and Optimization

DATA PROCESSING FOR ATOMIC ABSORPTION SPECTROMETRY WITH A MICROCOMPUTER

F. W. WILLMOTT and I. MACKENZIE*
Philips Research Laboratories, Redhill, Surrey, RH1 5HA (Gt. Britain)
(Received 12th May 1978)

SUMMARY

A general-purpose microcomputer has been designed and built to provide a data acquisition and control system. The microcomputer can be interfaced to an atomic absorption spectrometer to provide reliable and economic processing of data. The system provides scope for expansion in order to control various instrumental operating parameters simultaneously or consecutively.

The repetitive operations and complex calculations required in many forms of chemical analysis have created a demand for automation for both data handling and control purposes. This has been accelerated in recent years by advances in large-scale integration which have reduced both the cost and number of components required to perform a given function. Such advances are exemplified by the rapid developments in microprocessor technology. These devices are now appearing in a wide range of both utility instruments such as oscilloscopes and spectrum analysers and specialized analytical equipment including infrared and x-ray spectrometers and chromatographs. The functions which the microprocessors perform vary considerably from one situation to another. At one extreme, the device is simply used as an inexpensive substitute for hardwired logic requiring little of the potential speed and processing power. In other instances, it may be the central processing unit for a microcomputer and perform functions previously requiring a minicomputer. In the latter case, the microcomputer can offer significant advantages such as portability, relatively low cost and dedication to a particular instrument in situations where a minicomputer would not be economically viable.

Several different approaches to the application of a microprocessor in analytical work are possible. Some of the more powerful calculators can be programmed to acquire and analyse data from an instrument; this approach has been taken both by research groups [1] and several commercial manufacturers of analytical instruments such as atomic absorption spectrophotometers (a.a.s.). It offers the advantage of not dedicating a powerful tool to one technique, and allowing the user to specify program changes as desired. Disadvantages are the limitations on memory capacity, the lack of software support

facilities and the associated slowness of software development. A second approach is with a microcomputer dedicated to a particular technique such as the integrators used for data handling and control in chromatography [2]. These instruments implement relatively complex algorithms but offer no opportunities for the user to alter software to his own requirements other than by specification of a few parameters.

The third possible strategy is to apply a general-purpose microcomputer, possibly with a specific interface board, when the software can be exactly tailored to the requirements of the operator. Such a microcomputer can then either be dedicated to a particular analytical area or even an instrument, or used in a variety of alternative roles as the need arises [3]. Relatively long and complex algorithms can be implemented, because most 8-bit microcomputers have an expandable memory facility of up to 64 kbytes. Additional computer features, including a real-time clock, interrupts and input/output facilities, are also available with this approach as well as a range of software support facilities depending on the particular microprocessor. Disadvantages arise from the effort required to develop appropriate software and the possibl redundancy of some components. Software aids have greatly improved since the innovative days of microprocessors and development times can now be considerably reduced, especially with the aid of machine-resident or cross compilers for high-level languages such as PL/M, FORTRAN and CORAL 66.

This laboratory has used a number of general-purpose microcomputers with different analytical techniques for data capture and processing as well as for instrumental control The present paper describes how one of these computers was interfaced to a particular analytical instrument, an atomic absorption spectrometer, with emphasis on minimizing the required hardware and software effort.

EXPERIMENTAL

The microcomputer (Fig. 1) is designed around an Intel 8080A from a set of standard cards (Philips Science and Industry Group, Eindhoven). The basic computer consists of three cards incorporating the microprocessor, a universal asynchronous receiver/transmitter, memory and associated circuitry This is normally supplemented with further cards providing a real-time clock, interrupt facilities, a digital-to-analog converter and peripheral interfaces. The computer is completed with a 16-bit integrating analog-to-digital converter of the voltage-to-frequency type (Dynamic Measurements Corp. Type 8412) coupled with a variable-gain input amplifier. The voltage rails of +5V and ±12V are supplied from a Coutant EMP-15 unit. The total available memory consists of 12 K bytes and can be readily expanded as desired; this comprises 8 kbytes of random access memory (RAM) for storage of both variables and programs under development and 4 kbytes of u.v.-erasable programmable read-only memory (EPROM) containing standard input/output routines, a system monitor and mathematical routines including packages for curve fitting and floating point arithmetic.

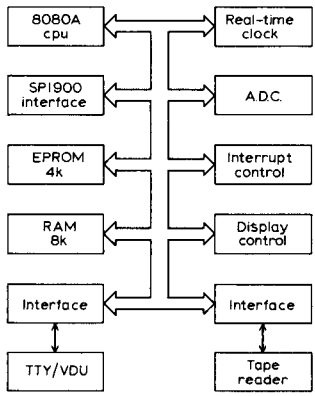


Fig. 1. Microcomputer architecture.

Communication to and from the computer requires either a video display unit (Lear Siegler ADM3A) or a teletypewriter (Texas Instruments, Silent 700). Object programs are either entered into RAM from paper tape by a high-speed tape reader or can be permanently stored in EPROM once fully developed. The former alternative is invariably used during program development stages and may be economically preferable even for debugged programs when these are used on an irregular basis; storage of an 8 kbyte program in 2708 EPROM currently requires some $220 of memory (U.K. prices, January 1978).

The data processing algorithm

The program is written in an interactive form by means of dialogue with the operator. Invalid replies by the operator are signalled as errors on the output device and the question is repeated. A flow chart is shown in Fig. 2 (a—e).

The data are sampled at 5 Hz for 20 s resulting in 100 readings which are stored as floating point values. The mean of these values is calculated and any "wild" data points greater than three standard deviations from the mean are rejected and a new mean is calculated. A calibration curve is constructed by fitting the readings from a series of standard solutions of known concentrations (not more than 15) by using a quadratic least-squares fitting procedure. The errors between the fitted line and the actual values are displayed so that the operator may reject any particular value if the deviation from the line appears excessive. The fit is then repeated on the remaining points. Results from up to 15 sample solutions are then calculated by fitting the data to the calibration curve. The concentration of the element is finally calculated and expressed relative to either the solution of the original sample in parts per million or as a percentage. A typical example of part of the microcomputer output is shown in Fig. 3.

Software was written to analyse data when the spectrometer was operating in the flame atomization mode. This could be readily adapted to the different

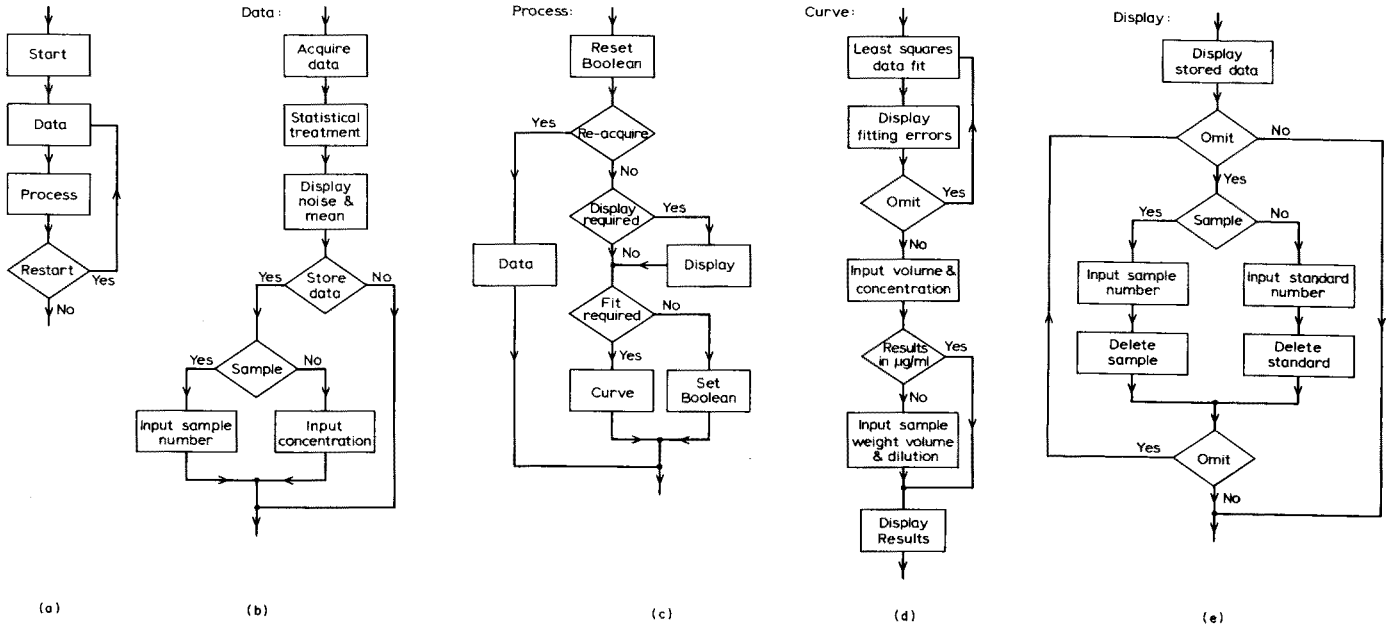


Fig. 2(a). Overall flowchart.
Fig. 2(b). Flowchart for subroutine "Data" used in Fig. 2(a).
Fig. 2(c). Flowchart for subroutine "Process" used in Fig. 2(a).
Fig. 2(d). Flowchart for subroutine "Curve" used in Fig. 2(c).
Fig. 2(e). Flowchart for subroutine "Display" used in Fig. 2(c).

START DATA ACQUISITION WITH A ONE _1
DATA ACQUISITION COMPLETE

MEAN VALUE 00401.15
REJECTED VALUES 00000
STD. DEV. 00004

STORE ? TYPE Y OR N _Y
STORE IN SAMPLE STORE ? TYPE Y OR N _N
CONCENTRATION ? TYPE VALUE _25

REACQUIRE ? TYPE Y OR N _N
DISPLAY ? TYPE Y OR N _Y

STANDARDS
MLS

00000	00000
00005	00093
00010	00180
00015	00259
00020	00334
00025	00401
00030	00466

SAMPLES
NUMBER

00001	00329
00002	00349

OMIT ? TYPE Y OR N _N
REACQUIRE ? TYPE Y OR N _N
DISPLAY ? TYPE Y OR N _N
FIT ? TYPE Y OR N _N

STD(MLS)	ERROR(MLS)
00000	– 00000.0375
00005	00000.0738
00010	00000.0114
00015	00000.0048
00020	– 00000.1083
00025	00000.0258
00030	00000.0375

OMIT ? TYPE Y OR N _N
STD.CONC.(μGS/ML) = _28.54
STD.VOL.(MLS) = _100.0
DO YOU REQUIRE CONCS IN SOLN? TYPE Y OR N _Y

SAMPLE DATA

SAMPLE 00001

CONCEN = 00005.63 PPM

SAMPLE 00002

CONCEN = 00006.04 PPM

Fig. 3. Typical computer output for an a.a.s. experiment. For space economy, the second part of the output is placed alongside the first part, instead of following on as would happen in practice.

requirements of flameless operation, i.e. faster signal acquisition and integratic of the peak area, in conjunction with the data analysis kit supplied by the manufacturer.

The spectrophotometer interface

The a.a.s. to which the microcomputer was interfaced was a Pye Unicam SP1900 double-beam instrument. In deciding on the link between the microcomputer and the SP1900, there were three main points to be considered: (i) there should be no alterations or additions made to the existing hardware inside the SP1900; (ii) the interface should be simple with the minimum of extra hardware; and (iii) neither the SP1900 nor the microcomputer should in any way be affected by the physical connection or the extra hardware. It should be possible to operate the two instruments independently whilst they are connected together.

The SP1900 provides detector signals in 3 forms: (i) a 10-mV analog signal; (ii) a 1-V analog signal; and (iii) a BCD signal updated at a rate of 5 Hz. The 10-mV signal is primarily for chart recorder applications with a bandwidth of 20—30 Hz and is hence unsuitable for this application. The 1-V signal is pulse-height modulated, making it incompatible with a voltage-to-frequency type ADC which expects a continuous signal over the conversion period. The BCD output signal after conversion back to TTL levels was therefore most suitable to transmit to the computer. This is normally used in conjunction with a Philips tally roll printer via an output socket and the interface board was designed with a plug to accept this socket directly. Timing of data transfer was achieved by edge triggering with an internal control signal available from the spectrophotometer.

Software development

Software was written in high-level language modules whenever possible by using a subset of the real-time language, CORAL 66, for which a cross compile is available to run on the laboratory computer, an ICL 1904S. It is well known that high-level languages produce inefficient code in terms of both execution speed and memory requirements but this factor is not critical in the present application. The only real-time processing of data required in the 200-ms intervals between acquisition of consecutive data points is conversion of an input from BCD to binary format, which requires approximately 500 μs.

Arithmetic definition of the parameters involved in the algorithm requires a combination of 8-bit, 16-bit and 24-bit integer or floating point arithmetic. The 8080 instruction set is most efficient when operations are limited to the first two options but certain stages of the algorithm, statistical analysis of the input signal, calculation of a least-squares fit and evaluation of sample concentrations or weights require floating point arithmetic to maintain adequate accuracy. The cross compiler used did not provide facilities for incorporation of floating point variables or calculations and these routines were inserted into the assembly language output from the compiler.

The overall program was cross-assembled on the ICL 1904S and the hexadecimal machine code output onto paper tape. This could be input to the microcomputer via a high-speed tape reader and subsequently debugged via the system monitor, which consists of a series of EPROM resident procedures providing facilities for the setting of breakpoints, substitution of memory contents, examination of the microprocessor register contents, etc.

The total memory requirements of a debugged program occupied 1.5 kbytes of EPROM and 5.8 kbytes of RAM. Software development took a total of 10 man weeks from definition of the project requirements to production of a fully debugged program. This period could be further reduced by addition of a development system based on floppy discs and an improved compiler for the 8080A incorporating floating point arithmetic; both have become available to users since this work was initiated.

RESULTS

This particular SP1900 is normally connected to a Honeywell 516 minicomputer as part of a distributed network for data analysis. Analytical results output by the minicomputer therefore provide a useful comparison for the microcomputer when the instruments are run simultaneously. Some typical figures quoted in Table 1 illustrate the final calculated concentrations of solutions of three metallic cations. Identical results are not obtained from the microcomputer and minicomputer since two different floating point packages are used operating on 16-bit and 24-bit mantissae, respectively.

Table 2 illustrates precision tests on raw spectrometric output data acquired by the two computers. Data acquisition was synchronized manually and produced the small differences in mean data values.

DISCUSSION AND CONCLUSIONS

This work illustrates how an inexpensive general-purpose microcomputer can be used in conjunction with an atomic absorption spectrometer for routine data analysis.

TABLE 1

Comparisons of results between microcomputer and a Honeywell 516

Element	Concentration in solution (ppm)	
	Microcomputer	Honeywell 516
Manganese	3.912	3.929
	3.240	3.255
Iron	5.638	5.630
	6.042	6.034
Copper	5.800	5.775
	8.878	8.867

TABLE 2

Precision tests with manganese solutions

	Microcomputer	Honeywell 516
Mean	750.6	750.8
Readings	20	20
S.d.	1.5	2.4
R.s.d. (%)	0.2	0.3

The modular construction of the software enables different algorithms to be evaluated or incorporated with relative ease into programs to analyse a particular problem. Storage of programs externally to the microcomputer facilitates its use as a general-purpose laboratory computer. For example, this particular instrument has been used to analyse very fast or transient input signals, control various instruments via transducers and analyse data from alternative forms of analytical instrumentation [4, 5]. The use of common softwa procedures greatly shortens development times in these instances.

The system has proved very convenient to use, providing fast sample throug put and offering considerable potential for elaboration. The software uses very little real-time processing since calculations are performed after the run. Hence there is more than sufficient processing power available for simultaneo control of instrumental operating parameters. Examples include switching power to the hollow-cathode lamp, setting the operating current, slit width, wavelength and integration periods. Addition of an automatic sampler and turntable under microcomputer control would then result in an automatic spectrometer.

The authors thank F. Grainger and I. Gale for advice on a.a.s., and W. Baig for interfacing the spectrophotometer and microcomputer.

REFERENCES

1 J. J. Pireaux, Appl. Spectrosc., 30 (1976) 219.
2 S. P. Cram, S. N. Chesler and A. C. Brown III, J. Chromatogr., 126 (1976) 279.
3 M. Goedert, S. A. Wise and R. S. Juvet, Chromatographia, 7 (1974) 539.
4 I. Mackenzie, W. G. Baig and F. W. Willmott, Proc. SERT Symposium, Microprocessor Systems and Software, University of Kent, 1977.
5 F. W. Willmott, I. Mackenzie and R. J. Dolphin, Proc. 12th Int. Symp. on Chromatography, Baden-Baden, J. Chromatogr., in press.

Analytica Chimica Acta, 103 (1978) 409—415
Computer Techniques and Optimization

THE APPLICATION OF LINEAR DISCRIMINANT ANALYSIS IN THE DIAGNOSIS OF THYROID DISEASES

D. COOMANS, M. JONCKHEER and D. L. MASSART*
Vrije Universiteit Brussel, Bosstraat, B-1090 Brussels (Belgium)
I. BROECKAERT
Sint Pieters Ziekenhuis, Internal Medicine Department, Brussels (Belgium)
and P. BLOCKX
A. Z. Middelheim, Nuclear Medicine Department, Antwerp (Belgium)
(Received 11th May 1978)

SUMMARY

The effectiveness of five in-vitro laboratory tests for differentiation between three thyroid functional states (EU, HYPO and HYPER thyroidism) has been determined by using statistical linear discriminant analysis. The optimal linear combination of laboratory tests obtained by means of linear discriminant analysis results in a better use of the information present in each test, so that the possible redundancy of tests can be assessed. In this context, some feature selection criteria were evaluated. It is shown that in this application only two laboratory tests are necessary to obtain a sufficiently high diagnostic effectiveness when linear discriminant analysis is applied.

Analytical chemistry is frequently concerned with the classification of individuals or samples into groups on the basis of chemical results. If these groups are pathological cases, the process of classifying is then called diagnosis. The evaluation of clinical laboratory tests according to their diagnostic effectiveness by using multivariate classification rules is therefore of interest in analytical chemistry. The methods used to do this are the same for medical diagnosis as for the more usual analytical applications, such as the classification of archeological artefacts or oil spills.

In the present investigation, five laboratory tests specific for thyroid diseases were evaluated by means of statistical linear discriminant analysis (s.l.d.a.). Usually these tests are processed as independent entities, although the physician implicitly combines them in making his diagnosis. Pattern recognition makes it possible to combine the test results explicitly. The purpose of this investigation was to use pattern recognition methods in medical diagnosis, and particularly to try to combine diagnostic effectiveness with a smaller number of tests. Pattern recognition specialists call this feature selection.

The dataset comprised 215 patients from the same hospital. These individuals were divided into 3 groups of known classification, namely: (a)

euthyroid patients (EU) (150 cases); (b) patients suffering from hyperthyroidism (HYPER) (35 cases); and (c) patients suffering from hypothyroidism (HYPO) (30 cases). Each individual was represented by a pattern vector comprising the results of five laboratory tests: total serum thyroxine (T4), total serum tri-iodothyronine (T3) or (T3RIA), T3 resin uptake (RT3U), serum thyroid-stimulating hormone (TSH), and increase of TSH after injection of TSH-releasing hormone (ΔTSH). The 3-fold classification problem (between EU, HYPER and HYPO) was divided into two binary sub-problems, namely HYPO/EU and HYPER/EU discrimination. The HYPO/HYPER discrimination was not studied because a 100% correct differentiation between HYPO and HYPER individuals is possible with a single test. Besides, physician order tests according to the expected pathology and a clinical impression of hyperthyroidism cannot be confused with a clinical impression of hypothyroidism.

Evaluation of the results

A question which arises in this context is how to present the results. For a chemist, it may be as bad to place a sample belonging to category A in category B as to place a category B sample in category A. For a physician, there is a distinct difference between considering an ill person as normal and a healthy person as sick. In a screening procedure, it is preferable to classify a few normals provisionally as ill than to classify ill people as normal. Therefore a somewhat more complex evaluation system is necessary. Four categorie should be considered [1]:

(a) true positives, i.e. persons positive to the test indicating illness who are indeed ill (TP);

(b) true negatives, i.e. persons negative to the test who are indeed normal (TN

(c) false positives, i.e. persons positive to the test who are in fact normal (FP);

(d) false negatives, i.e. persons negative to the test who are in fact ill (FN).

In the context of pattern recognition "positive to the test" and "negative to the test" should be replaced by "classified as positive (ill)" and "classified as negative (normal)".

The following definitions are then used:

sensitivity = [TP/TP + FN] $\times$ 100 (% of ill people detected).

selectivity = [TN/(TN + FP)] $\times$ 100 (% of normals identified as such).

and the combined expression:

efficiency = [(TP + TN)/(TP + TN + FP + FN)] $\times$ 100 (% correctly classifie

Youden index = sensitivity + selectivity $-$ 100.

CLASSIFICATION WITH 5 LABORATORY TESTS

In the first instance the diagnostic effectiveness of statistical linear discriminant analysis (s.l.d.a.) was determined by means of the 5 original laborato

tests. The computations were done by using the commercially available Statistical Package for the Social Sciences (SPSS) [2]. Table 1 lists the results for the HYPO/EU and HYPER/EU classification. It can be seen that for both discriminations the efficiency and Youden index are influenced only by the selectivity of the classification procedure, because a sensitivity of 100% is obtained. Transformation (log or square root) of the data does not improve the results and is therefore not considered further here. The result obtained with 5 tests is acceptable but not excellent. Since it was thought that this could be due to "noise" added by one or more of the tests, feature reduction was attempted.

Feature selection

Five laboratory tests, i.e. 5 features, were used. To establish if these 5 tests are really necessary and if the diagnosis can be made with fewer tests, the essential features which must be retained have to be recognised. Feature selection is perhaps the most interesting aspect of pattern recognition in medical research: while few physicians will currently accept that classification by pattern recognition is a worthwhile diagnostic tool, some physicians at least accept that mathematical methods make it possible to select or develop economically sound, better sets of laboratory tests for particular diagnostic problems. For this reason, and because little information is available in the analytical literature about it, feature selection by parametric methods is discussed here in some detail.

Feature selection is done on the basis of selection criteria, which attach numerical values to the discriminatory ability of each laboratory test individually or in relation to others. In this investigation, only parametric criteria were used, the numerical values of which are based on statistical considerations. There are two main reasons why laboratory tests can be eliminated: either because they contain little information, i.e. alone the test gives little discrimination between the pathological groups studied, or because they contain correlated information, i.e. the test in question gives little additional information compared to another test.

Typical selection criteria are available for both kinds of redundant features. Selection criteria that trace only tests containing little information are further referred to as criteria of the first kind, and criteria which also take the correlation into account are referred to as criteria of the second kind. Criteria of

TABLE 1

Diagnostic effectiveness of s.l.d.a. for 5 tests

Discrimination	TP% (Sensitivity)	TN% (Selectivity)	EFF% (Efficiency)	YI% (Youden index)
HYPO/EU	80.0	100.0	96.6	80.0
HYPER/EU	91.4	100.0	98.4	91.4

the first kind are, for example, variance ratio, Fisher weighting [3], and resolution. These attach numerical values to the information content of each test separately, without consideration of correlation with other tests. As an example, only the resolution was evaluated for the present data set. The resolution between a normal (EU) group and a pathological (P) group for laboratory test i, may be stated as

$$R_i = (|\overline{X}_{P,i} - \overline{X}_{EU,i}|)/(S_{P,i} + S_{EU,i}) \quad (1)$$

where $\overline{X}_{p,i}$ and $\overline{X}_{EU,i}$ are the mean values of test i for groups P and EU, respec tively, and $S_{p,i}$ and $S_{EU,i}$ are the standard deviation of test i for groups P and EU, respectively. Greatest importance is attached to the laboratory test with the highest R_i value.

Selection criteria of the second kind may be divided into direct and stepwise procedures. In the direct procedure, the absolute value $|v_i|$ of the discriminant weight coefficient and the percentage contribution to the discrimination (C) were used as criteria. The absolute value of the discriminant weight coefficient for test i is defined [4] as

$$|v_i| = \sum_{j=1}^{n} |g_{j,i}(\overline{X}_{P,j} - \overline{X}_{EU,j})| \quad (2)$$

where $g_{j,i}$ is element (j,i) of the inverse pooled variance—covariance matrix and n is the total number of laboratory tests. The percentage contribution to the discrimination can be expressed as

$$C = (|v_i\,\delta_i|/D^2) \times 100$$

where $\delta_i = \overline{X}_{P,i} - \overline{X}_{EU,i}$ and $D^2 = \Sigma_{i=1}^{n} |v_i\delta_i|$, n indicating the total number of tests. For each criterion, the best test is the one with the largest $|v_i|$ or C value.

In the stepwise procedure, only combinations of laboratory tests are considered. Initially, the single test which has the best value for the selection criterion is chosen, in the present case by Rao's V statistic [2]. This initial test is then paired sequentially with each of the other available tests and V is computed again. The test which produces the best value of V in conjunction with the initial test is selected as the second test. Computation is continued in this way until all variables are included. The first test to be included can be considered as the most important one and the last as the least important.

Results of feature reduction

In Table 2, the results obtained by selection criteria are compared for the HYPO/EU discrimination. Each column represents the sequence of decreasin importance according to a given selection criterion. The same sequences are obtained for the $|v_i|$ and C_i selections and also for the R_i and the stepwise selections. All criteria agree that the two most important tests are T4 and ΔTSH. The agreement between the selection criteria of the first kind and those of

TABLE 2

Comparison of selection criteria for the HYPO/EU discrimination

R_i	$\|v_i\|$	C_i	Stepwise
T4	T4	T4	T4
ΔTSH	ΔTSH	ΔTSH	ΔTSH
RT3U	TSH	TSH	RT3U
TSH	T3RIA	T3RIA	TSH
T3RIA	RT3U	RT3U	T3RIA

the second kind for these two tests is probably due to the small importance of the correlation between these tests and the rest. Table 3 gives the within-groups correlation matrix for the 5 tests.

Table 4 gives the results of the classification after feature selection. S.l.d.a. was applied after each deletion of the least important test. Clearly, the selectivity increases significantly with decreasing number of laboratory tests, and RT3U, TSH and T3RIA do not contribute to the discrimination. In fact, they cause confusion; in other words, they only add noise. The sensitivity is always 100%.

Table 5 lists the results for the HYPER/EU discrimination. The selection criteria are compared in the same way as for the HYPO/EU discrimination, and the agreement is clearly less successful. Only the first test (T4) corresponds for all 4 criteria. Only the C_i and the R_i selection agree for the 3 most important tests.

TABLE 3

Within-groups correlation matrix for the HYPO/EU discrimination

r	RT3U	T4	T3RIA	TSH	ΔTSH
RT3U	1.000				
T4	0.169	1.000			
T3RIA	0.184	0.311	1.000		
TSH	0.083	−0.200	−0.186	1.000	
ΔTSH	−0.005	0.077	0.148	0.097	1.000

TABLE 4

Selectivity for the HYPO/EU discrimination after stepwise selection

Number	Selectivity
5	80.0
4	83.0
3	86.7
2	96.7

TABLE 5

Comparison of selection criteria for the HYPER/EU discrimination

R_i	$\|v_i\|$	C_i	Stepwise
T4	T4	T4	T4
ΔTSH	T3RIA	ΔTSH	RT3U
T3RIA	RT3U	T3RIA	TSH
RT3U	ΔTSH	TSH	T3RIA
TSH	TSH	RT3U	ΔTSH

Table 6 shows the classification after each reduction step for each selection criterion. In contrast to the reduction for the HYPO/EU discrimination, a small, but continuous and therefore significant decrease of the diagnostic effectiveness is observed after each reduction step, i.e. no single test is really redundant. The T4—ΔTSH combination again scores the highest Youden index and highest efficiency although the difference is too small to be considered significant.

VALIDATION OF THE CLASSIFICATION WITH TWO LABORATORY TESTS

The study described above indicates that the T4—ΔTSH combination is the most appropriate for both the classifications HYPER/EU and HYPO/EU. This conclusion was reached by feature selection on the complete data set. It remains to be shown that the classification results obtained are also valid when the classification rules are applied to new cases. To do this, a "leave-one-out" procedure was used; this means that one of the cases was eliminated,

TABLE 6

Feature selection for the HYPER/EU discrimination

Criterion	Number of variables used	Sensitivity	Selectivity	Efficiency	Youden index
R_i	5	100.0	91.4	98.4	91.4
	4	100.0	91.4	98.4	91.4
	3	99.3	88.6	97.3	87.9
	2	97.3	88.6	95.7	85.9
$\|v_i\|$	4	100.0	91.4	98.4	91.4
	3	97.1	96.7	96.8	93.8
	2	82.9	98.7	95.7	81.6
C_i	4	100.0	91.4	98.4	91.4
	3	99.3	88.6	97.3	87.9
	2	97.3	88.6	95.7	85.9
Stepwise	4	100.0	88.6	97.8	88.6
	3	98.7	85.7	96.2	84.4
	2	98.7	77.1	94.6	75.8

TABLE 7

Prediction performances estimated by the "leave-one-out" procedure

	Selectivity	Youden index
HYPO/EU	96.7	96.0
HYPER/EU	88.6	85.9

the classification rules were determined on the remaining data, and classification of the eliminated case was then studied. This was repeated for each of the cases. The percentage of cases correctly classified then gives an estimate of the prediction performance. Table 7 shows the results.

The values in Table 7 are very similar to those obtained during the feature selection procedure (see Table 4 for the HYPO/EU classification and Table 6 for the HYPER/EU classification). This validates the results obtained, and it can be concluded that three of the five laboratory tests investigated (RT3U, TSH, T3) are completely redundant — even confusing — for the HYPO/EU classification and largely redundant for the HYPER/EU classification.

REFERENCES

1 R. S. Galen and S. R. Gambino, Beyond Normality, J. Wiley, New York, 1975.
2 N. H. Nil, C. H. Hull, J. G. Jenkins, K. Steinbrenner and B. H. Bent, Statistical Package for the Social Sciences, McGraw-Hill, New York, 1975.
3 B. R. Kowalski, Chemometrics, Theory and Application, ACS Symposium Service 52, American Chemical Society, Washington, 1977.
4 M. Kendall, Multivariate Analysis, Charles Griffin & Co, London, 1975.

Analytica Chimica Acta, 103 (1978) 417—428
Computer Techniques and Optimization

EVALUATION OF THE SUPER-MODIFIED SIMPLEX FOR USE IN CHEMICAL PATTERN RECOGNITION

STEVEN L. KABERLINE and CHARLES L. WILKINS*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588 (U.S.A.)

(Received 3rd May 1978)

SUMMARY

A simplex optimization technique, the super-modified simplex (SMS), is evaluated for use in the pattern recognition analysis of low-resolution mass spectra. For the recognition of eleven functional group categories, the performances of SMS-derived weight vectors are shown to be comparable to those obtained by a previously developed modified simplex method. Data are presented which indicate that the SMS procedure requires fewer simplices and decreased computational time to converge to an optimized solution for the structural analysis problems investigated.

Recent studies of chemical pattern recognition have demonstrated the usefulness of the method for application to spectral analysis [1, 2]. As part of continuing research devoted to the development of a collection of classfiers to serve as a significant diagnostic aid in spectroscopy, various pattern recognition techniques have been investigated [3—5]. A primary objective of the work presented here was critically to evaluate the performance of a new optimization procedure, the super-modified simplex (SMS), and its applicability to pattern recognition analysis of low-resolution mass spectra.

The linear learning machine (LLM) method of pattern recognition is an empirical approach to data analysis [6]. In its application to spectra, training sets of data obtained from compounds of known functional group composition are used to train weight vectors to make binary decisions. Figure 1

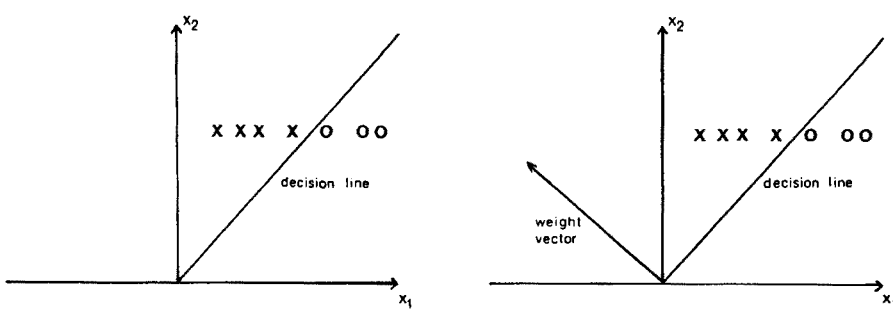


Fig. 1. Two-dimensional illustration of the relationship between pattern space and weight vector space.

depicts this process for a two-dimensional example. For the present application, mass spectra are represented as points in a space of sufficient dimensionality. Accordingly, the X and O symbols in this example each represent one mass spectrum with two peaks. It is assumed that compounds having similar structures will have similar mass spectra and will, therefore, cluster within the same region of space. The linear learning machine starts with an arbitrarily chosen decision line (or hyperplane, for cases involving more than two dimensions), and then "moves" this line, by using an error-correction feedback process, in an attempt to find the decision surface that perfectly divides the data into the desired two categories. Ultimately, the process yields a weight vector which can then be applied to an unknown pattern to allow a decision to be made regarding the presence in the compound of the functional group which the weight vector is trained to recognize. The learning method of developing linear binary pattern classifiers has been investigated extensively, primarily because of the relative computational ease of development and application of such classifiers. The algorithm will always converge to a solution if the data are linearly separable (i.e., if a hyperplane does exist that can perfectly divide the data into two categories), but the rate of convergence cannot be predicted. A serious disadvantage of the method is the fact that it cannot deal adequately with the linearly inseparable data presented by most actual chemical structural problems. If the data are linearly inseparable, then no perfect weight vector solution exists, and the weight vector fluctuates randomly in the vector space formed by the weights. Stated simply, there is no way for the LLM to find an optimal weight vector solution if the data are linearly inseparable.

THE MODIFIED SIMPLEX METHOD

The simplex method is an optimization process first proposed by Spendly et al. [7] and later modified by Nelder and Mead [8]. The simplex approach to optimization has been successfully applied to a variety of problems in analytical chemistry [9—11] and has been recently adapted for use in pattern recognition studies [12]. Modified simplex (MS) pattern recognition is capable of producing near optimal weight vectors even when inseparable data sets are used for training [3].

A simplex is a geometric figure having $d+1$ vertices in a d dimensional space. For example, a simplex in two dimensions is a triangle, and a three-dimensional simplex is a tetrahedron. Figure 2 illustrates a simplex in a two-dimensional space. For the present applications, the simplex operates in response space, not pattern space or feature space as does the linear learning machine (see Fig. 1). Each vertex of the simplex is a weight vector, and associated with each weight vector is a response. Here, the response being optimized is recognition, i.e. the number of training set patterns correctly classified. A second optimization criterion, the "perceptron" criterion, is also used in order to insure that the response space is continuous. The perceptron

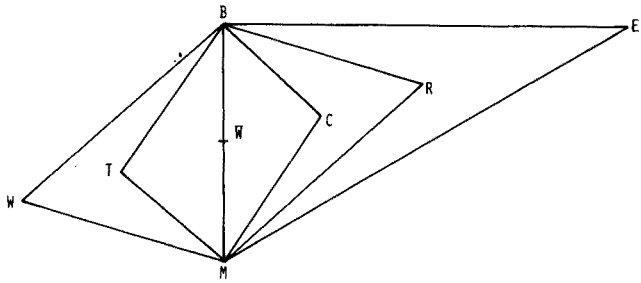


Fig. 2. Vertices involved in movement for the modified simplex method.

is defined here as the sum of the absolute values of the dot products of all misclassified patterns with the weight vector, and is only used as a smoothing function, when it is not possible to determine the best response from the recognition values alone. The best vertex of the simplex is, therefore, the one with minimum perceptron for the weight vector giving maximum recognition. The starting simplex is defined by assigning a judiciously chosen value to one of the initial weight vectors. The other vertices must be chosen so that weight vector space is spanned. This is done by forming the remaining initial weight vectors by adding an appropriate "spanning constant" to each of the components (weights) of the first weight vector [12]. The object of simplex optimization in pattern recognition is to find the maximum response (defined as above) by moving the simplex. Movement is done by replacing methodically and iteratively the worst vertex with a better vertex. The simplex thus moves toward the optimum in weight vector space. For example, in Fig. 2, if W represents the vertex with the worst response (poorest recognition), B the vertex with the best response (highest recognition, or if the recognition is the same as at W, lowest perceptron), and M an intermediate response, W will be reflected through the centroid ($\bar{W}$) of the remaining vertices to form the new vertex R. Now, if R is better than B, correct movement is indicated and the simplex is further expanded to E. The alogorithm incorporates several other such strategies [12] intended to allow the simplex to proceed to the optimum with maximum speed and efficiency. The movement of the simplex continues until an optimum has been sufficiently closely defined. If the data are found to be separable, which is usually not the case for real chemical structural analysis problems, the search of weight vector space will terminate at 100% recognition with a perceptron of zero. If the data are not separable, the simplex will be halted when it appears to be stranded in a small region of space; that is, it will halt after several successive iterations indicate no improvement in recognition, and no significant decrease in perceptron criterion. Obviously, other criteria could also be used to indicate that the simplex has indeed reached an optimum solution [12].

Extensive use of simplex pattern recognition has shown that MS-developed weight vectors are generally superior in their performance to the corresponding LLM weight vectors, particularly if the data are linearly inseparable. There are, unfortunately, two disadvantages inherent in the use of the simplex

algorithm as it is applied to pattern recognition. First, there is no guarantee that the solution is the optimal solution. Response space usually contains a number of local maxima, and it is possible that a local, rather than a global, maximum has been reached. Secondly, the computational time involved can be prohibitive. A typical case (for the present research) can require more than an hour of IBM 370-148 CPU time to converge to an optimal solution. This is necessary because the method requires the calculation of a response at least once for every iteration, and finding a response involves computing the dot product of a weight vector (vertex) with every pattern in the training set.

THE SUPER-MODIFIED SIMPLEX METHOD

The super-modified simplex (SMS) method was recently developed by Denton et al. [13], and used for the on-line optimization of experimental parameters (fuel and oxidant flow, burner heights, etc.) in computer-controlled flame spectrometry. Their results showed that, compared to the modified simplex, the SMS algorithm required an average of 72% fewer simplices, 12% fewer data points, and an average saving of 21% in required time to converge to an optimum solution.

The SMS procedure applied to pattern recognition analysis of chemical data is quite similar to the MS method. Both methods use the same procedure to define the initial, starting simplex. In addition, the MS and SMS techniques both optimize the recognition for a given training set, and use the same criterion for defining when an optimum has been reached. For the present application, the only difference is that the SMS uses a different algorithm for moving the simplex towards the optimum. The modified simplex always moves in a precisely fixed manner; it may be expanded or contracted to allow it to follow more closely the contours of the response surface, but still is always moved by a constant amount.

Figure 3 illustrates how the super-modified simplex moves in weight vector space. For each iteration, the responses at the worst vertex (W) of the simplex, the centroid ($\bar{W}$) of the remaining vertices, and the point which is the reflection of the worst vertex through the centroid (R) are calculated. A second order polynomial curve is fitted to these three points. Furthermore, the curve is extrapolated beyond W and R by a fixed percentage of the W to R distance, resulting in two distinct curve types.

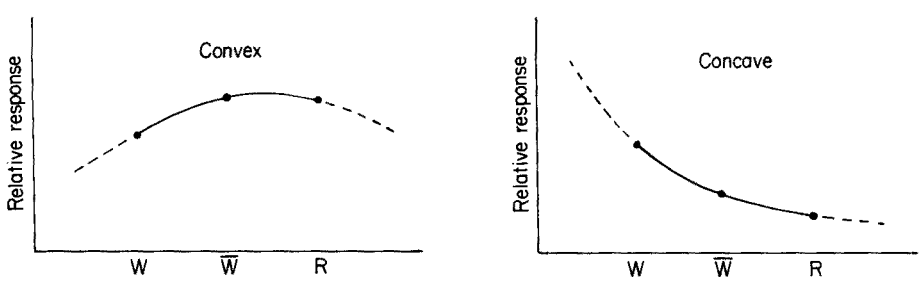


Fig. 3. Possible response contours for the super modified simplex.

For the convex case, a maximum exists somewhere (not necessarily within the interval W to R). Evaluation of the derivative of this curve allows finding the precise location of the predicted optimum. At this point, three distinct situations may arise. First, the location of the predicted optimum is within the region W $-x$(R $-$ W) to R + x(R—W), where x is the extrapolation factor. For this case, the response at the predicted optimum is calculated, the new vertex is located at this predicted point, and the optimization process continues. Secondly, the predicted maximum is not within the region W to R (including the extrapolation distance). Here the maximum within the region is taken to be the predicted maximum, and the new point will be assigned this position in space. Finally, the predicted optimum could lie very close to the centroid of the simplex. Clearly, if the new vertex was located at the centroid, the dimensionality of the new simplex would be reduced, and a search of certain portions of space would be prohibited. A small "safety" factor is always used to locate the new vertex a percentage of the W to R distance away from the centroid and prevent such a move.

For the concave case, no maximum occurs within the interval. The SMS algorithm selects the interval boundary producing the highest response as the new vertex. Again, the response at this predicted point will be calculated, a new worst vertex will be found, and the entire process will be repeated for the next iteration.

For both the concave and convex cases, if the predicted point does not have a better response than the reflected point, the reflected point is, instead, chosen as the new vertex. The SMS method, unlike the MS procedure, allows the new vertex to be assigned any of an infinite number of positions (excluding a small region around the centroid) along the line joining W and R bounded by the extrapolation limits. This ability to locate the position of the new vertex more precisely than is possible by the MS algorithm means that the super-modified simplex requires fewer iterations to converge to an optimized solution. This is its great strength in experimental optimization situations. Moreover, for pattern recognition applications (unlike optimization of instrument parameters), there is no need to consider boundary condition violations — by definition, it is impossible to have greater than 100% or less than 0% recognition.

EXPERIMENTAL

Data

For training and testing, a set of 1252 low-resolution mass spectra was drawn from a computer-readable file [14] of 18,806 spectra. These spectra were selected so that each compound represented contained only one of eleven functional groups chosen for the development of weight vectors (an exception to this was the phenol category, which obviously contains a substituted phenyl). For each of the eleven categories (Table 1), a training set of up to 210 spectra were chosen from the 1252 spectrum set. Whenever

TABLE 1

Composition of low-resolution mass spectra data

Category	Compound class	Number of spectra
1	Arenes	249
2	Aldehydes and ketones	96
3	Ethers	103
4	Aliphatic alcohols	185
5	Phenols	84
6	Carboxylic acids	51
7	Thiols	135
8	Esters	125
9	Amines	131
10	Amides	56
11	Nitriles	37

possible, training sets comprised approximately equal numbers of spectra of class members and non-members. The remaining spectra, unused in training, were used as test sets to determine the prediction capabilities of the weight vectors developed. Four smaller, two-category, two-dimensional training sets were also generated in order to test further the performances of both simplex algorithms in the absence of complications arising from the nature of mass spectral data. Two of the training sets were composed of 16 patterns each; the remaining two were composed of 20 patterns each. One of the 16 pattern and one of the 20 pattern training sets were linearly separable; the others were known to be inseparable. After weight vectors were developed from these training sets, a random number generator was used to generate a test set of 1000 class and 1000 non-class patterns.

Computations

Programs for modified sequential simplex and supermodified simplex optimizations, as well as a prediction program used to tabulate the performances of the developed weight vectors as applied to each test set, were written in FORTRAN IV, and all computations were performed with an IBM 370/148 computer. All simplex calculations for each training set were limited to a maximum of 20 min of CPU time. The spanning constant, used in developing the initial, starting simplex, was 4000 for all categories, except amines, where it was 100. The extrapolation factor used in SMS calculations was 50% of the W—R distance, and the safety factor was 5% of this same distance.

Preprocessing and feature selections

Intensities in the master file were encoded to 1 part in 9999, with the base peak assigned the value 9999, and the other intensities assigned values relative to the base peak. Metastable peaks and those with less than 1% relative intensity were not used. The square roots of the remaining intensities were used for subsequent calculations. Feature selection, i.e., the choice of

the m/e values to be used for development of the weight vectors, was accomplished by selecting the 60 peaks containing the most information, according to recent studies and calculations using information theory [15].

RESULTS AND DISCUSSION

Table 1 shows the distribution of compound types for the entire set of 1252 spectra which served as the source of the training and test sets. These compounds are monofunctional (except, of course, for categories 1 and 5). The categories were chosen to ensure that the training sets would be comprised of compounds with spectra uncomplicated by the presence of more than one kind of functional group (again, with the exception of the phenyl and phenol classes). The training sets used to develop the MS weight vectors were identical to those used in developing SMS discriminant functions, and both methods were restricted to the same fixed CPU time limit. Clearly, any differences in the performances of the weight vectors developed must be due to a genuine difference in the performance of the algorithms themselves, and not due to a difference in the difficulty of the problems presented to each simplex technique.

Tables 2 and 3 summarize the recognition and prediction performances for all MS and SMS weight vectors developed in this study. Recognition of training set members was, with one exception, slightly better for the modified simplex classifiers. However, eight out of the total eleven super-modified simplex classifiers showed better predictive ability than the corresponding MS-developed weight vectors. The differences were, in each case, small—generally less than 10% — which suggests that, perhaps, there is no significant difference in classifier performance for the structural problems investigated. Indeed, one of the problems associated with comparing classifiers solely on the basis of overall correct prediction is that this criterion can be shown to be significantly dependent on the test set composition. Since one of the main objectives of this study was to determine which simplex technique is capable of producing the more reliable classifiers, for cases where test set composition is unknown, it is particularly important to use an objective measure of classifier performance that minimizes test set dependence.

An alternative measure of classifier performance [16], derived from the earlier work of Rotter and Varmuza [17], can be based on recognition of the fact that the maximum possible information gain of a classifier is limited by the initial uncertainty, which in turn depends on the composition of the test set. This measure, which may be called the figure of merit, M, is the information gain after application of the classifier relative to the maximum possible information gain imposed by the test set composition. Tests with reasonable ranges of test set composition have shown that M, as a measure of classifier performance, does not suffer from the defect of extreme test set dependence. M is a dimensionless quantity, ranging from a value of 0 (no information gain) to a value of 1 (the classifier has yielded maximum possible information gain).

TABLE 2

Results for 60-feature MS recognition and prediction[a] expressed as percent correct monofunctional unknowns[b]

Weight vector	1	2	3	4	5	6	7	8	9	10	11
Recognition	99.0	98.0	92.0	90.5	98.0	92.0	98.5	92.4	81.7	95.5	100.0
Class membership/ prediction											
1	91.3	89.9	96.3	94.4	66.4	89.0	81.0	92.7	63.5	97.1	94.9
2	98.8	90.0	85.5	82.1	90.2	100.0	86.1	81.4	90.5	95.1	87.7
3	98.9	82.8	76.7	43.5	96.6	79.6	80.4	46.2	92.3	94.3	97.7
4	92.3	85.5	63.9	89.4	93.3	87.5	83.5	70.1	95.8	90.6	97.0
5	34.2	91.3	97.2	94.9	85.0	87.0	78.4	98.7	64.9	100.0	97.1
6	81.4	90.7	85.4	34.9	87.8	52.4	53.5	44.2	83.7	88.9	100.0
7	97.6	91.7	80.0	74.4	83.2	80.0	94.6	88.1	38.1	90.0	99.2
8	90.5	83.9	75.9	50.9	93.8	87.3	73.5	91.1	94.6	96.4	94.6
9	99.2	81.0	89.7	83.2	99.1	94.8	84.8	95.8	91.8	40.0	100.0
10	100.0	72.9	82.6	56.0	100.0	80.4	85.7	71.4	18.4	88.5	100.0
11	90.9	84.9	96.8	78.8	81.5	71.9	84.4	87.5	90.6	100.0	88.2
Average	88.7	85.9	84.6	71.1	88.8	82.7	80.5	78.8	74.9	89.1	96.0

[a]MS and SMS (Table 3) training and test sets were identical.
[b]Weight vector 1 was developed to recognize the presence of the Class 1 functional group, etc. Class numbering corresponds to that in Table 1. Each column of the table contains the detailed performance of the class weight vector specified for unknowns drawn from each of the classes. To test each weight vector, approximately 1000 unknowns were used.

TABLE 3

Results for 60-feature SMS recognition and prediction[a] expressed as percent correct monofunctional unknowns[b]

Weight vector	1	2	3	4	5	6	7	8	9	10	11
Recognition	98.5	97.0	90.0	90.0	96.5	91.5	95.5	89.5	87.1	92.5	97.5
Class membership/ prediction											
1	94.6	92.7	95.4	94.8	63.1	96.2	88.2	94.5	83.1	98.1	95.8
2	98.8	90.0	97.6	82.1	98.8	100.0	93.0	87.2	90.5	93.8	88.9
3	92.4	86.2	67.4	41.3	98.9	84.1	72.8	55.0	93.4	92.1	94.3
4	93.4	88.6	73.4	96.5	98.8	91.9	88.8	80.8	97.0	92.4	95.8
5	35.5	89.9	97.2	93.6	85.0	97.1	71.6	97.3	91.9	100.0	91.3
6	93.0	90.7	85.4	32.6	100.0	47.6	72.1	48.8	88.4	77.8	100.0
7	97.6	90.8	77.5	74.4	89.1	93.3	83.6	83.9	91.5	93.3	94.2
8	94.0	86.6	83.0	45.7	99.1	92.8	60.2	80.0	96.4	85.5	98.2
9	99.2	85.4	98.3	77.3	100.0	96.6	99.2	98.3	91.8	50.0	100.0
10	100.0	91.7	97.8	50.0	100.0	97.8	100.0	83.7	18.4	80.8	100.0
11	81.8	81.8	96.8	36.4	77.8	68.8	84.4	93.8	90.7	96.3	70.6
Average	89.1	88.6	88.2	65.9	91.9	87.8	83.1	82.1	84.8	87.3	93.6

[a,b]See footnotes to Table 2.

When the M values listed in Table 4 are compared, it can be seen that, except for Classes 9 and 11, there are no significant differences in the performance of the weight vectors developed by MS or SMS. It is interesting to note that even though the figures of merit are quite similar, indicating nearly the same level of performance for the corresponding weight vectors, the weight vectors themselves are quite different. Comparing for example, the components (weights) of the MS weight vector for phenyls to the same SMS weight vector, there is often a remarkable difference. The corresponding components of the weight vectors are not only different in magnitude, there is frequently a difference in sign. This further demonstrates the fact that weight vector space may contain several maxima, since one would expect nearly identical final weight vectors if the same maxima in space had been reached by movement of the simplex. It appears that these two algorithms, for reasons not entirely clear at this time, usually converge to maxima located in different regions of response space, even though the resulting final weight vectors differ little in their performance with the test sets used in this study.

Another measurement of classifier performance is tabulated in Table 5. These are the a posteriori probabilities, $P(1/J)$ and $P(2/N)$. $P(1/J)$ is the probability that a prediction of Class 1 membership will be correct and $P(2/N)$ that a prediction of Class 2 membership will be correct. These quantities obviously depend on the test set composition, but, nevertheless, represent a quantity of particular interest to the user of pattern classifiers. They answer the useful question of "given that the classifier predicts a particular class membership for an unknown, what is the expectation that the prediction is correct?" For both the MS and SMS methods, if a classifier predicts that a particular functional group is absent from a compound represented by a mass spectra, there is indeed a very high probability that that

TABLE 4

Comparison of test results for MS and SMS methods[a]

Class	Figure of merit	
	MS	SMS
1	0.49	0.54
2	0.33	0.37
3	0.20	0.19
4	0.24	0.28
5	0.29	0.33
6	0.09	0.12
7	0.32	0.26
8	0.28	0.24
9	0.24	0.39
10	0.32	0.26
11	0.48	0.30

[a]These values were computed from the data used in compiling Tables 2 and 3.

TABLE 5

Prediction results (1052 compound test sets)

Class	$P(1/J)$		$P(2/N)$	
	MS	SMS	MS	SMS
1	0.60	0.62	0.98	0.99
2	0.17	0.24	0.99	0.99
3	0.24	0.20	0.99	0.98
4	0.20	0.23	0.99	0.99
5	0.08	0.25	0.99	0.99
6	0.21	0.13	0.99	0.99
7	0.18	0.22	0.99	0.99
8	0.18	0.19	0.99	0.99
9	0.16	0.31	0.99	0.99
10	0.16	0.15	0.99	0.99
11	0.29	0.21	0.99	0.99

functional group is actually not present. The a posteriori probability $P(1/J)$ for a positive classification, i.e., the classifier does predict the presence of the functional group, shows more of a difference between the classification abilities of MS- and SMS-developed classifiers. Both of these a posteriori probabilities are involved in the figure of merit calculations, and it is primarily the differences in $P(1/J)$ probabilities that account for the differences in M. Thus $P(1/J)$ is the quantity most important to the user of the pattern recognition classifiers. Knowing that a compound does not contain a specific functional group is of relatively little value compared to knowing that a given functional group is present in the unknown compound.

It is now worthwhile to consider the differences in the two algorithms in terms of speed and efficiency. A good approximation of how "fast" each of the algorithms is can be obtained by examining how many iterations each method was able to complete during the fixed CPU time limit imposed on all of the optimization runs. The modified simplex was able to find an optimized solution for only one category (amines) during the 20-min time limit. For each of the rest of the categories, the MS program performed an average of 2541 iterations during this time. The supermodified simplex algorithm, however, was able to converge to an optimized solution in less than 20 min for five (4, 7, 8, 9, and 11) of the eleven categories. For category 9, an optimum amine weight vector was calculated in only 42 iterations. An average of 1176 iterations was required for the remaining six classes. It can be seen that the SMS is, indeed, considerably slower (about half as fast) per iteration. However, it usually requires far fewer iterations (about half as many) to obtain a weight vector of comparable or slightly better performance when compared to the MS-derived vectors. A detailed examination of the algorithm logic reveals the source of the differences in speeds of the two methods. The fitting of a quadratic curve involves relatively trivial mathe-

matical calculations, and this contributes little to the slow computational speed of the super-modified simplex for the present application. Rather, it is the measurement of a response at a vertex itself that is the limiting factor for algorithm speed. Calculating a response, in these studies, required computing the dot product of a 60-component weight vector with each pattern in a training set of 200 compounds, a process which requires a time several orders of magnitude greater than fitting three points to a quadratic curve. The super-modified simplex always requires the finding of three responses per iteration: the response at the centroid, the reflected point, and the new predicted optimum. The MS, however, requires finding the response a variable number of times per iteration. The response is calculated at least once (for the reflected point), and at most twice (depending on what contractions or expansions are required, following examination of the response at the reflected point). Tabulating the number of times the response must be calculated for a typical modified simplex run shows that, on average, the modified simplex requires finding a response slightly less than 1.5 times per iteration. This, then, indicates why the SMS requires approximately twice as long per iteration. The SMS does, however, appear to be more efficient, for it still converges to a comparable level of performance after only half as many iterations. In other words, the SMS moves directly toward an optimum; the MS apparently "wanders" somewhat during its moves towards a response surface maximum.

As a final comparison between the modified and super-modified simplex techniques, small, two-dimensional training sets were artificially generated. Data sets 1 and 2 were separable, while sets 3 and 4 were known to be inseparable. Table 6 summarizes results obtained from these data. Again, the same trend of CPU time and number of iterations is seen for these examples; The SMS algorithm is slower per iteration but, for the more difficult problems (category 3, for example), requires fewer iterations. The last column in Table 6 represents the performance of weight vectors developed on a randomly generated test set consisting of 1000 class and 1000 non-class members. In each case, the SMS-produced weight vector shows equivalent or slightly better performance for the classification of the "unknowns". As in the cases involving "real" chemical data, there is usually a marked difference in the values of the components of the weight vectors developed by the two methods.

It can be concluded that there is a slight advantage in using the SMS method for the development of optimal or near-optimal weight vectors for the recognition of important functional groups. The SMS algorithm requires about twice the time per iteration compared to the MS technique, but the algorithm does move more efficiently towards an optimum, even for cases involving linearly inseparable data. For pattern recognition responses, the SMS and MS algorithms usually approach optima located in different regions of the response space; this may result in classifiers which can provide additional useful information regarding the presence of a functional group for a particular application.

TABLE 6

Comparison of modified and super-modified simplex methods with artificial 2-dimensional data

Data set[a]	Recognition training set		CPU time (s)		Iterations[b]		Classification number correct[c]	
	MS	SMS	MS	SMS	MS	SMS	MS	SMS
1	16	16	0.77	1.81	5	10	1497	1560
2	20	20	0.85	1.87	5	10	1384	1711
3	14	14	1.27	2.41	112	84	1141	1002
4	18	18	1.39	2.70	107	106	1028	1099

[a]Sets 1 and 2 contained 16 and 20 patterns and were separable. Sets 3 and 4 contained 16 and 20 patterns and were inseparable.
[b]A minimum of 10 iterations was required with the algorithm employed for SMS to terminate the calculation.
[c]Test sets were random two-dimensional data and contained 2000 members.

It is evident that although the dramatic improvements realized by Denton and co-workers [13] in experiment optimization applications are not realized for pattern recognition studies, nevertheless, there is a demonstrable advantage in the use of the technique.

The authors gratefully acknowledge the valuable critical suggestions contributed by Dr. R. B. Spencer, and the financial support by the National Science Foundation under grant CHE-76-21295. A grant from the University of Nebraska Research Council provided partial support for the purchase of the mass spectral data collection used.

REFERENCES

1 T. L. Isenhour, B. R. Kowalski, and P. C. Jurs, Crit. Rev. Anal. Chem., July, 1974, pp. 1—44.
2 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, J. Wiley, New York, 1975.
3 T. F. Lam, C. L. Wilkins, T. R. Brunner, L. J. Soltzberg, and S. L. Kaberline, Anal. Chem., 48 (1976) 1768.
4 L. J. Soltzberg, C. L. Wilkins, S. L. Kaberline, T. F. Lam, and T. R. Brunner, J. Am. Chem. Soc., 98 (1976) 7143.
5 C. L. Wilkins and T. R. Brunner, Anal. Chem., 49 (1977) 2136.
6 N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
7 W. Spendly, G. R. Hext, and F. R. Himsworth, Technometrics, 4 (1962) 441.
8 J. A. Nelder and R. Mead, Comput. J., 7 (1965) 308.
9 R. R. Ernst, Rev. Sci. Instrum., 39 (1968) 998.
10 D. E. Long, Anal. Chim. Acta, 46 (1974) 1170.
11 S. N. Deming and S. L. Morgan, Anal. Chem., 46 (1974) 1170.
12 G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, Anal. Chem., 47 (1975) 1951.
13 M. Wm. Routh, P. A. Swartz, and M. B. Denton, Anal. Chem., 49 (1977) 1422.
14 E. Stenhagen, S. Abrahamssen, and F. W. McLafferty, The Registry of Mass Spectral Data, J. Wiley, New York, 1974.
15 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
16 L. J. Soltzberg, C. L. Wilkins, S. L. Kaberline, T. F. Lam, and T. R. Brunner, J. Am. Chem. Soc., 98 (1976) 7139.
17 H. Rotter and K. Varmuza, Org. Mass Spectrom., 10 (1975) 874.

Analytica Chimica Acta, 103 (1978) 429—443
Computer Techniques and Optimization

FOUR LEVELS OF PATTERN RECOGNITION

CHRISTER ALBANO, WILLIAM DUNN III, ULF EDLUND, ERIK JOHANSSON, BO NORDÉN, MICHAEL SJÖSTRÖM and SVANTE WOLD*

Research Group for Chemometrics, Umeå University, S-901 87 Umeå (Sweden)

(Received 3rd May 1978)

SUMMARY

Problems of pattern recognition in chemistry and other subjects can be divided conveniently into four different types depending on the level of scope of the problem.
(1) Classification into one of a number of defined classes. As an example blood samples taken from persons known to be either controls or welders are considered. The problem is whether trace element concentrations in these samples contain information on whether or not a person is a welder.
(2) Level 1 plus the possibility that an object is an outlier, i.e. does not belong to any of the defined classes. As an example, the use of ^{13}C-n.m.r. data to decide whether 2-substituted norbornanes have the *exo* or *endo* structure is discussed. (2A) Level 2, asymmetric. This situation occurs when one class does not have a systematic structure, but another class is homogeneous and can be described by a level 2 model. This occurs in the classification of materials or compounds as good or bad, active or inactive, and in binary classifications. As an example the use of trace element data to classify steel samples as having good or poor properties of strength is discussed.
(3) Level 2 plus the ability to relate the variables measured to external properties of continuous character. As an example, the classification of a series of chemical compounds as β-receptor blockers, β-receptor stimulants, or neither, on the basis of their structural variables is discussed. In addition, relations between these structural variables and the measured biological activity are sought within each of the two classes.
(4) Level 3 with the difference that several external property variables in the objects are measured. It may be desirable to use variables of the objects both for classification and for relations to several property variables: such examples are numerous in analytical chemistry.

Multivariate data, i.e. data sets where each object is characterized by several variables, are increasingly common in analytical chemistry. Thus the analysis of organic samples by gas chromatography, liquid chromatography, electrophoretic methods, gel filtration, etc., often gives for each sample 5—100 "peaks", the heights or areas of which are related to concentrations of sample components. Trace element analysis is used to characterize samples in terms of "concentrations" of 10—30 trace elements. Spectroscopic methods give spectra which can be translated into a large number of variables, e.g. position and height of characteristic peaks, coupling constants, etc.

To relate multivariate data sets to the chemical problems involved, methods of pattern recognition are becoming popular [1—4]. Several methods of

pattern recognition have been applied to chemical problems, most frequently the linear learning machine (LLM) [1—4], linear discriminant analysis (LDA) [5, 6], the K nearest-neighbor method (KNN) [1—4] and, more recently, the SIMCA method [7, 8].

Pattern recognition is often formulated as a methodology for finding rules of classification, i.e. given a number of classes, each of which is defined by a set of objects (the training or reference sets) and the values of M measurements made on each of these objects, rules are defined that make it possible to classify new objects (the test set) on the basis of the same M measurement made on these new objects. In analytical chemistry, the objects are usually samples of some kind, i.e. mixtures of chemical compounds or sometimes a pure chemical compound for which the structure is to be determined. The variables — the measured data — are obtained by chemical analysis as mentioned above. Finally, the classes are chosen to correspond to the analytical problem. In a structural analysis of a compound, the classes involved are structural types, i.e. alcohols, amines, hydrocarbons, etc. On other occasions, it may be desirable to determine the source or the type of a particular sample e.g. an archaeological artefact. The classes then correspond to the possible different sources or types and the reference sets (training sets) of each class contain samples that are known to have their origin in the corresponding source.

In chemistry, the scope of a data analysis is usually wider than merely obtaining a classification of unassigned objects. Usually, it cannot be certain that an object does not belong to a yet unseen class. It is necessary to be able to discriminate not only among the known classes, but also among possible unknown classes. It is also desirable to obtain an empirical description of each class to allow for predictions and correlations with external property variables. It is convenient to order pattern recognition problems in a hierarch depending on the scope of the data analysis. This allows for an understandin of the applicability of pattern recognition to various kinds of chemical problems. Also, it becomes easier to choose an appropriate method of pattern recognition for the scope of the analysis of the particular problem.

A graphical representation of pattern recognition

The M data observed for one object can be represented as a vector of lengt M, represented in turn as a point in the M-dimensional space obtained by giving each variable one coordinate axis. Since M is usually larger than 3, this M-space is difficult to draw on paper. However, three-dimensional M-spaces can be used as illustrations with the conjecture that higher-dimensional space have analogous properties.

In M-space a class of objects is represented as a swarm of points and severa classes as several, distinct or overlapping, swarms. Pattern recognition can be seen as a methodology to describe the class swarms quantitatively so that it is possible to calculate to which of the class swarms the point of a new objec is closer (Fig. 1).

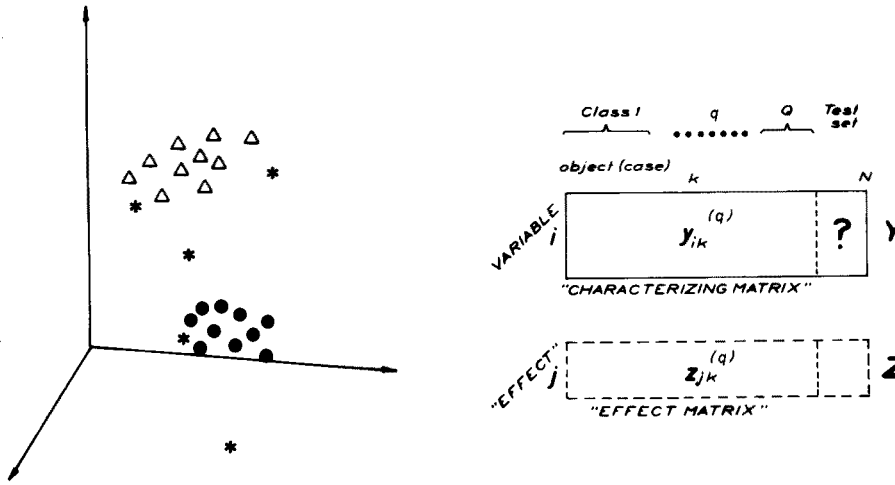


Fig. 1. Three-dimensional measurement space ($M = 3$) with objects from two classes (triangles and circles) and some unassigned objects (asterisks).

Fig. 2. Data available in pattern recognition problems. The matrix Z is available only in level 3 and 4 problems.

FOUR LEVELS OF PATTERN RECOGNITION

Before going into the different levels of the problem, the data available in pattern recognition problems are summarized (Fig. 2). The observed value of variable i for object k in the reference set of class q is denoted by $y_{ik}^{(q)}$. The number of variables is M, the number of objects in the qth reference set is n_q and the number of classes is Q. Further, there are objects in the test set with data $y_{ik}^{(?)}$. Sometimes these are also values of external property variables, $z_{jk}^{(q)}$, to be related to the characterizing variables $y_{ik}^{(q)}$, e.g. in the prediction of physical, chemical or biological properties of chemical compounds. For instance, data Y for a number of compounds may be required to classify compounds as active or inactive against, say, a certain bacterial strain, and to obtain a detailed prediction of the level of activity for drugs based on a relation between the measured activity of some of the compounds in the reference set and their corresponding characterizing data Y.

In principle, it is assumed that all positions in the data matrix Y (and Z) are defined by actual observed values for objects in the training set. In practice, this is sometimes not the case. This missing observation problem can be handled in various ways, many particular to a method of pattern recognition as discussed below.

Level 1. Classification into either of a number of defined classes

On this lowest level it is assumed that all objects in the training and the test sets belong to one of the initially defined classes. As an example, Table 1

TABLE 1

Number of cases in the training set of five classes in the example of the welders (data from ref. 9). The number of cases correctly classified by linear discriminant analysis (LDA) and SIMCA are also given.

			LDA	SIMCA
Class 1	Controls	$n_1 = 68$	26	27
2	type a	$n_2 = 23$	10	9
3	b	$n_3 = 7$	6	4
4	c	$n_4 = 28$	13	14
5	d	$n_5 = 23$	8	8
$Q = 5$		$N = 149$	63	62

shows data for blood samples from controls (class 1) and from welders involved in four different methods of welding (class 2—5); the concentrations of 17 trace elements were determined in each blood sample. The problem was to find out how to use these 17 variables to determine if a sample originated from a normal individual (class 1) or a welder affected by the working environment (class 2—5). Details of the analysis have been discussed [9].

A pattern recognition problem on level 1 can be solved in several ways; most methods operate on, and are designed for, this level. If the reference sets of the classes can be separated from each other by some surfaces, new objects can be classified according to the side of these planes on which they fall (Fig. 3). This is the principle of the linear learning machine (LLM) and linear discriminant analysis (LDA) — the separating surfaces are planes computed to achieve maximal separation of the classes. The calculation of these planes corresponds mathematically to the computation of linear combinations of the variables (discriminant functions) which have different signs for objects situated on different sides of the planes.

In the K nearest-neighbor (KNN) method, an object in the test set is classified according to the class of its K nearest neighbors in M-space. Usually K is taken as 1 or 3 (Fig. 4). SIMCA can also be made to work on this level, but is described below since it operates naturally on level 2.

The information obtained on level 1 contains two parts.

(a) Class separation. When the objects in the training set are classified according to the derived rules, a crude measure of the class separation is obtained. In the welders example, LDA classified the objects in the training set as shown in Table 1, indicating a poor class separation. This measure of class separation is over-optimistic because the same data set is used both for the calculation and the testing of the classification rule [10, 11].

(b) Classification of the test set. The objects in the test set are classified according to the derived rules. A lower limit of the error rate of classification is obtained from the classification of the training set — about 43% in the welders example. No test set is present in this example. Therefore the only information is the somewhat discouraging result that the available data give

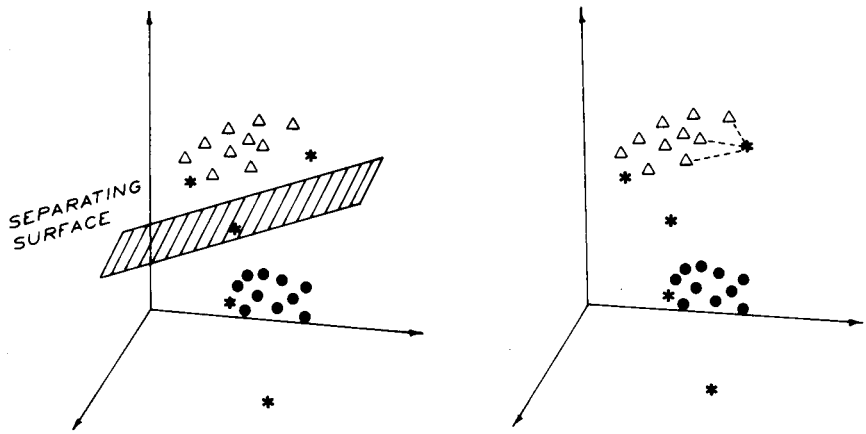


Fig. 3. LDA and the LLM operate by means of surfaces separating the classes in M-space.

Fig. 4. The KNN method classifies an unassigned object according to the class of its K nearest objects in M-space (usually K is one or three).

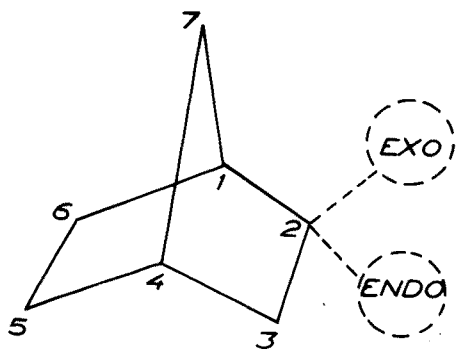
ca. 43% separation of the five classes. This is certainly too low to give a basis for medical decisions about future blood samples — one rather wonders if the data contain any information at all about the working environment of welders.

Level 2. Level 1 plus the possibility of outliers

The assumption made on level 1, i.e. that all objects belong to one of the defined classes, is often unrealistic. Thus, objects in the reference sets might have been selected erroneously or might otherwise be atypical for the class. An object in the test set might be of a new type, e.g. a member of a hitherto unknown class.

As the first data set on this level 2, the ^{13}C-n.m.r. data [12] of a number of *exo* and *endo* 2-substituted norbornanes (Fig. 5) are considered. The n.m.r. shift of each of the seven carbons (1—7 in Fig. 5) were measured for 8 *exo* compounds and 7 *endo* compounds in the hope that the data would be useful in structural analysis of other 2-substituted norbornanes. In this data set, it is not absolutely certain that the assignments of the spectra are correct for all compounds. Moreover, some compounds included in the test set differ somewhat in structure from those in the training set; some have carbonyl groups at position 5 or 6, others have a double bond between C-5 and C-6 and still others have two methyl groups in position 7. Hence, the possibility of outliers must be taken into account. The data set on welders is also considered to illustrate the additional information obtained on this level 2.

On level 2 pattern recognition works by containing each class in a closed mathematical structure — an envelope — in M-space. The class envelopes are



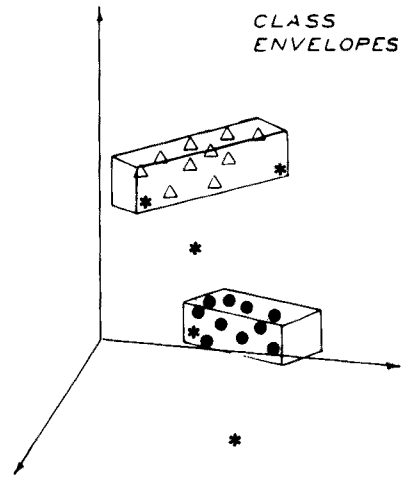


Fig. 5. Chemical structure and numbering of 2-substituted norbornanes.

Fig. 6. On level 2, each class reference set is contained in a closed structure (envelope). In this figure, the closed structures are the parallelepipedes of the entropy minimax method (EMM).

constructed so that an object inside an envelope is considered to be a normal member of a class and objects outside all envelopes are outliers to all classes (Fig. 6). Two methods operate on this second level, namely the entropy minimax method (EMM, [13, 14]) and the SIMCA method (simple modelling of class analogy [7, 8]).

In the EMM, each class is enclosed in one or several parallelepipedes which are constructed so as to minimize the disorder (entropy) of each class (Fig. 6). In the SIMCA method, an A-dimensional hyper-plane is fitted to each class. A confidence "slab" is constructed around the plane on the basis of the distribution of the objects with respect to the plane. In addition the slab is limited along the plane on the basis of the distribution of the objects along the corresponding dimensions. This gives the closed class envelopes shown in Fig. 7.

The concept of closed class envelopes not only allows for the detection of outliers, but also provides other valuable information, e.g. measures of the relevance of each variable, and measures of inter-class distances. Two ways are seen to measure the relevance of a variable. The modelling power relates to the extent of participation of a variable in description of the classes. It is related to the within-class variation of a variable compared with its total variation over the whole training set; i.e., the average thickness of the envelopes along the variable coordinate axis compared with the total range of the variable. The discriminating power of a variable relates to the extent of participation of a variable in discrimination between the classes. This is measured as the average distance between the class envelopes along the coordinate axis

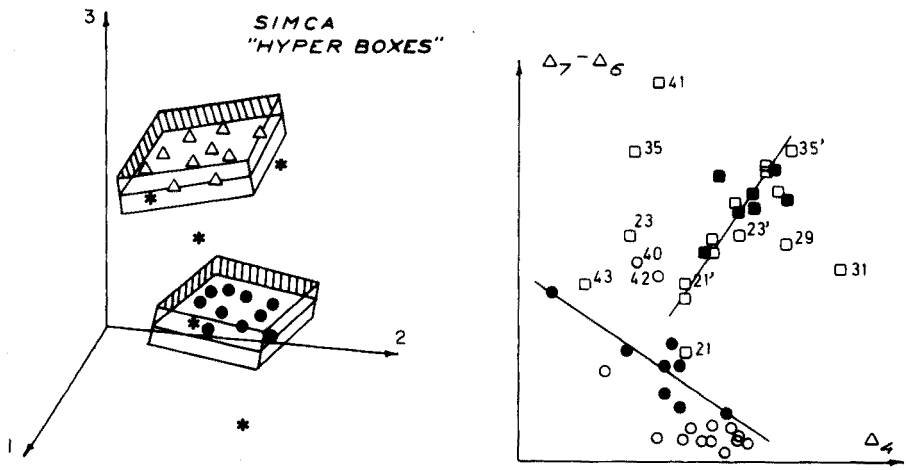


Fig. 7. In the SIMCA method, the class envelopes are constructed from the planes best fitting the objects in each reference set and confidence regions around this plane. Variable 3 has good modelling power because the thickness of the class envelopes along the vertical axis is small compared with the data range along the same axis. The opposite is true for variable 2. The discriminatory power of the variables in this illustration parallels the modelling power, but this is not always the case. The distance between the two classes, finally, is about 5 as seen from the separation of the "hyper boxes" compared with their thickness.

Fig. 8. A projection of the three most discriminating variables in the ^{13}C-n.m.r. example (norbornanes) onto a plane. Here Δ_i denotes the relative shift of the i-th carbon, i.e. $\Delta\delta_{Ci}$. •○, *exo*; ▪□, *endo*. Filled symbols indicate objects in the reference sets; open symbols indicate objects in the test set. The primed numbers show the site of the objects with corresponding unprimed numbers after reassignment of the n.m.r. spectra. Objects 29, 31 and 40—43 are outliers.

of the variable compared with the average envelope thickness along the same axis. The inter-class distance, finally, is measured as the distance between two envelopes relative to their average thickness, the measures being taken along a line connecting the two centers of the classes. Illustrations are given in Fig. 7 (see legend).

A SIMCA analysis of the ^{13}C-n.m.r. data reveals that only three of the seven variables contain discriminatory power, i.e. are essential for discriminating between *exo* and *endo* compounds. A projection of these variables is shown in Fig. 8. One clearly sees the two class structures and the good fit of most of the compounds in the test set to either of the two class models. Some of the compounds in the test set, however, are clear outliers (21, 23, 29, 31, 35 and 40—43). Of these, 21, 23 and 35 were shown to have an erroneous spectral assignment. When this was corrected, these three compounds fitted the *endo* class model well (points 21′, 23′ and 35′ in Fig. 8). (The remaining outliers (29, 31, 40—43) could not be made "normal" by spectral reassignment; it is concluded that they are not similar to the compounds in the training set.

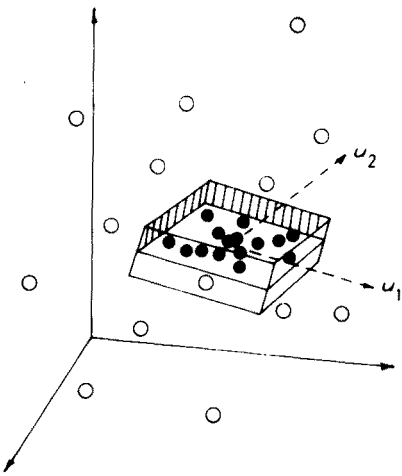


Fig. 9. The asymmetric case of pattern recognition. One class (•) forms a well-defined class that can be well described by a closed mathemetical structure, a SIMCA box. The other class (○) are spread more or less randomly in M-space and cannot be contained in a meaningful closed structure. The classification of new objects can still be made, however, by observing if they fall inside or outside the "class box", leading to a classification as class 1 (inside) or 2 (outside), respectively.

A partial explanation for this dissimilarity is the presence of two methyl groups in the 7-position in compounds 40—43, which might introduce new steric effects.

When the welders' data set is analysed on level 2, the most valuable information concerns the relevance of the variables and the inter-class distances. Of the 17 variables originally included in the analysis, 5 had low modelling and discriminating power; i.e. they lacked relevance for the specified problem of distinguishing between controls and welders and between different types of welders. When the data were re-analyzed with the remaining 12 relevant variables, the inter-class distance matrix obtained (Table 2) showed that there

TABLE 2

Distances between the 5 classes reference sets in the welders example. D_{ij} is the residual standard deviation obtained when the data in class i are fitted to the class model of class j. The elements D_{ij} are normalized by dividing by the residual standard deviation (s) of class j shown in column 7.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	s
$i = 1$	1.00	1.10	1.41	1.04	1.05	0.862
2	0.94	1.00	1.28	0.98	0.97	0.771
3	1.01	1.09	1.00	1.05	0.96	0.753
4	1.03	1.13	1.37	1.00	1.06	0.839
5	1.01	1.12	1.30	1.05	1.00	0.822

is absolutely no separation of the classes. Thus, even though there is some systematic structure in the data set of each class reference set, these structures are almost identical for all 5 classes. The conclusion is that the data contain absolutely no information about the specified problem.

Level 2A. Level 2, asymmetric

In the level 2 analysis above, it is assumed that all class reference sets are contained in closed envelopes. This corresponds to the assumption that each class contains similar objects, i.e. is homogeneous. In the analysis of data where the classification corresponds to a qualitative difference in "effect", this assumption of homogeneity is often not valid. Another situation, where one class is inhomogeneous, is encountered in so-called binary classification where one well-defined alternative is considered against all others, e.g. it is desirable to discriminate between alcohols and all other types of compounds on the basis of i.r. spectra.

As an example, an analysis of steel samples made for a Swedish steel company is discussed. The data set consists of 17 trace element concentrations measured on 15 steel samples; 6 of these (class 1) had little tendency to crack but 9 cracked (class 2) in a later stage of the manufacturing process. When these data were analyzed by SIMCA, the objects in class 1, the "good steels", had a nice data structure, well-contained in a closed envelope of small size. The "bad" objects in class 2 could not, however, be enclosed in any well-defined envelope; they seemed to be spread randomly in M-space (Fig. 9). The situation seems rather natural: "good" steel samples must meet rather stringent specifications, i.e. they must be very similar objects which can be contained in a small area of M-space; in contrast bad steel samples can result from many different manufacturing faults, and so can be expected to appear in an apparently random fashion; the class is inhomogeneous and without detectable structure.

This asymmetric situation, where one class has no detectable structure has also been found in the classification of chemical compounds as carcinogens or non-carcinogens on the basis of their chemical structure [15]; the carcinogens formed a good homogeneous class whereas inactive molecules showed no systematic behaviour in M-space.

On level 2 of pattern recognition, the problem of data analysis can still be solved. Since the "good steel" class is contained in a closed envelope, this can be used to classify future objects as "good" or "bad" according to whether or not they fall inside the "good" envelope. The further away from the envelope, the worse the quality of the corresponding object; 14 of the 15 steel samples were classified correctly by an "asymmetric" SIMCA analysis.

Level 3. Level 2 plus a prediction of one "external" effect variable z

Often the scope of the data analysis is two-fold: (i) to devise a classification on the basis of characteristic variables (y); (ii) to predict the level of an "external property variable", z, i.e. relate the level of z for objects in each

class to the variables y. In the steel example, it may be desirable to use trace element concentrations for a classification as discussed above, and to relate a quantitative measure of the "goodness" of the steels, say the corrosivity, to the same variables. Such data were not available in the steel example; instead, the methodology is illustrated with an example from chemical structure–biological activity relations [16].

Compounds of the general structure shown in Fig. 10 were tested for their β-receptor activity [17]. Of the 37 compounds, 17 were antagonists (blocking activity, class 1); 18 had a stimulating activity (activator, class 2); and 5 of the compounds had little or no activity. A quantitative measure of the activity, log C, was also measured for the 32 active compounds. In the data analysis, the chemical structure of each compound was described by substituent parameters and other physicochemical variables deduced from the structure, 13 variables in all. With these variables (y), a SIMCA analysis resulted in the description of the two classes by two separate "hyper-boxes" (Fig. 7). Of the 32 compounds, 30 were classified correctly as antagonist or activator with these hyper-boxes.

To approach the second goal, i.e. to relate the variables y to the measured level of activity (z) in each class, the assumption is made that just as objects from the same class are closely situated to each other in M-space, objects having a similar level of activity, z, are very close to each other in the same space. This corresponds to the assumption that a direction ζ in the envelope of a class can be found such that the position along the ζ-axis is correlated to the level of the external variable z.

SIMCA describes each class by means of a "fenced-in" hyper-plane (of A dimensions) which takes the mathematical form

$$y_{ik}^{(q)} = m_i^{(q)} + \sum_{a=1}^{A} b_{ia}^{(q)} u_{ak}^{(q)} + e_{ik}^{(q)}$$

Here, the parameters m_i and b_{ia}, estimated from the class reference set, define the position and direction of the hyper-plane. The coefficients u_{ak}, specific for the k-th object in the class, describe where the object is situated (see Fig. 9). The residuals, e_{ik}, give measures of how far the object lies from the plane.

Since the parameters u_{ak} relate directly to the position of the object within the class, the second part of the data analysis simply seeks a relation between the "effect" variable z and the coefficients u_{ak}. This is a simple

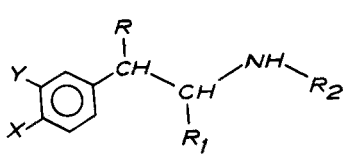


Fig. 10. Basic structure of the β-adrenergic compounds investigated. The substituents X, Y, R, R_1 and R_2 were described by 2 or 3 variables each, leading to a description of each compound in terms of 13 variables.

multiple regression of z on the A u-vectors. It is preferable, however, to have a graphical representation of the same analysis (Fig. 11) where the direction u_1—u_2 correlates remarkably well with the measured activity z (log C) of the activators, i.e. class 2. The antagonist class is not included but shows an equally good picture. The variable z is not included in the data analysis until the last stage; i.e. the u-vectors are calculated only to describe the common behaviour within each class. Thereafter, this remarkable correlation with the measured activity is a strong indication of the relevance of the u-vectors and the class structures.

Thus, the type and level of a new compound can be predicted. First, the chemical structure is translated into descriptor variables (found in tables). Secondly, on the basis of these variables, the compound is classified as activator or antagonist (or neither) on the basis of the fit of the data-vector (y) of the object to the two class models, i.e. whether or not the point of the object falls inside one of the two class hyper-boxes in M-space. Third, the resulting u values of the object are entered into the plot of the corresponding class, leading to the prediction of the level of the activity. Figure 11 shows such predictions for three test compounds known to be inactive or to have very low activity. Indeed, a very low activity is predicted.

Level 4. Level 3, but with several external property variables

A level 4 problem involves prediction of the level of several property variables (contained in the matrix Z) in addition to a classification. This is probably common in practice — in the steel example it may be desired to correlate a number of "goodness" variables, e.g. corrosivity, strength, hardness, etc., in addition to a classification such as "good" or "bad". In structure—activity relationships there are often several effect variables; in addition to the biological activity measured in a number of different biological test systems, there are negative side-effects such as toxicity, etc. Curiously, the level 4 problem

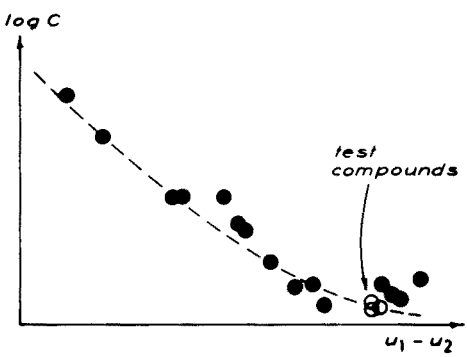


Fig. 11. Relation between measured activity (log C) of the 15 promoters and their position in the class structure as expressed by their class coordinates u_1 and u_2.

is easier, in principle, than the level 3 problem because the presence of several effect variables stabilizes the systematic structure of the effects in the same way as the presence of several characterizing variables stabilized the class structure in M-space of the y-variables.

The data analysis consists of two phases, as in level 3 problems. First, the systematic class structure is described in terms of closed envelopes in two separate spaces: one "characterizing space", defined by the y-variables, and one "effect space", defined by the z-variables. Irrelevant variables are deleted and the analysis is made as if it consisted of two separate level 2 analyses. In the second phase, a direction in each class in the Y-space that is correlated maximally to a direction in the corresponding class structure in Z-space is sought (Fig. 12). Mathematically, this corresponds to a multiple regression (if the envelopes in Z-space are one-dimensional), a canonical correlation [18] or, optimally, a NPLS analysis [19] for each class.

Though the possible applications of this level 4 analysis are numerous, the level 4 concept is quite new and a completed example is not yet available.

DISCUSSION

Chemical applications of pattern recognition can, in a rather natural way, be ordered into four categories; these four ambition levels and the mathematical methods involved have been discussed with a graphical exposition based on three-dimensional measurement spaces. The arguments derived from this graphical exposition can be translated into a stringent mathematical formalism for the benefit of those who require a rigorous treatment. The information available on levels 1—4 is summarized in Table 3.

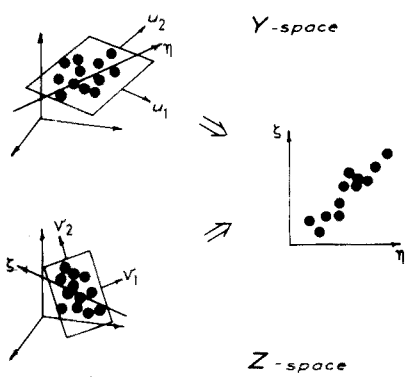


Fig. 12. Outline of a level 4 analysis by SIMCA. A closed structure of a class is defined in each of the two spaces corresponding to the characteristic variables (Y-space) and the effect variables (Z-space). A relation is then sought between a direction in the class model in Y-space and a direction in the class model in Z-space.

TABLE 3

Summary of information obtained on levels 1—4
(The level stated is the lowest level at which the information is available)

A. Information concerning objects	*B. Information concerning classes*
Level 1 Closest class	Level 1 Separation of classes
Level 2 Similar class—sufficiently close to indicate similarity between object and class	Level 2 Distances between classes
	Level 4 Distances between classes in terms of effect variables
Level 3, 4 Predicted value(s) of effect variable(s)	Level 1 Discrimination power of variables
	Level 2 Modelling power (explanatory power) of variables
	Level 4 Modelling power of effect variables
	Level 2 Data profile of class
	Level 3 Relation between effect and characterizing variables
	Level 4 Data profile of effect variables in class
	Level 2 Outliers in class reference set
	Level 4 Outliers according to effect variables

Philosophically, it is interesting that methods on level 1 are aimed at finding either differences between classes (LLM, LDA) or similarities between objects and classes (KNN) or within classes (Bayes). On level 2, the methods (EMM and SIMCA) are aimed at finding similarities within classes, i.e. finding closed structures containing each class.

Thus, with reference to the increased amount of information obtained at level 2, it seems advantageous to formulate problems in terms of similarity rather than in terms of dissimilarity. This can also be understood mathematically. The behaviour of a group (class) of similar phenomena can usually be treated with Taylor expansions and perturbation theory; similarity can rather easily be modelled mathematically [8]. Dissimilarity, however, is more difficult to handle in the same straightforward way; disparity usually cannot be Taylor-expanded by means of a series with a small number of terms.

By specifying the information desired from the data analysis, the ambition level and thereby the range of methods available for data analysis are specified. Most chemical data subjected to a pattern recognition analysis are analyzed at present on level 1. However, information such as the existence of outliers, distances between classes and the relevance of included variables can be obtained only on level 2 and higher. In fact, most pattern recognition problems in chemistry are at least level 2 problems. It is usually not sufficient to establish whether a new object (sample, compound or reaction) is more similar to this or that class. Rather, the desired information is whether the object is sufficiently similar to one of the classes to be considered a normal member of that class.

The asymmetric variant of level 2 classification is probably a rather common problem in chemical practice, particularly in so-called binary classification problems. Here spectra are used to determine whether or not a compound is of a given type, e.g. an aromatic hydrocarbon; i.e. a well-specified alternative against all other possibilities. In other instances, this binary classification is used to determine whether or not a sample comes from a given source, or is of a given type. The class of "all other possibilities" usually does not form a proper homogeneous group. Therefore, the problem usually cannot be treated as a level 1 problem. This is one reason for the apparent failure of pattern recognition analysis in several investigations involving binary classifications. The problem has been treated improperly because the asymmetry has not been realized.

The example of the welders shows that negative results are more conclusive on level 2 than on level 1. In practice, results of this type, i.e. a given data set which does not contain any information with respect to a given question, are valuable. At least, it is then certain that other types of data must be collected and analyzed to investigate the given problem.

When the variables included in a pattern recognition analysis have at least a large minority of "relevant"variables among themselves, however, mathematical methods find any information available, if they are correctly chosen and used. With the increasing frequency of multivariate problems in analytical chemistry, pattern recognition becomes a valuable tool, necessary for maintaining good research economy.

Support from the Swedish Natural Science Research Council is gratefully acknowledged. Helpful discussions with Herman Wold and Maynarhs da Koven have been of great value.

REFERENCES

1 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.
2 B. R. Kowalski, in C. E. Klopfenstein and C. L. Wilkins (Eds.), Computers in Chemical and Biochemical Research, Vol. 2; Academic Press, New York, 1974.
3 B. R. Kowalski, Anal. Chem., 47 (1975) 1152 A.
4 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley-Interscience, New York, 1975.
5 P. A. Lachenbruch, Discriminant Analysis, Hafner Press, New York, 1975.
6 T. Cacoullos (Ed.), Discriminant Analysis and Applications, Academic Press, New York, 1973.
7 S. Wold and M. Sjöström, in B. R. Kowalski (Ed.), Chemometrics, Theory and Application ACS Symposium Series, No. 52, Am. Chem. Soc., 1977.
8 S. Wold, Pattern Recognition, 8 (1976) 127.
9 U. Ulfvarsson and S. Wold, Scand. J. Work, Environ. Health, 3 (1977) 183.
10 N. A. B. Gray, Anal. Chem., 48 (1976) 2265.
11 C. P. Weisel and J. L. Fasching, Anal. Chem., 49 (1977) 2116.
12 M. Sjöström and U. Edlund, J. Mag. Reson., 25 (1977) 285.
13 R. Christensen, Entropy Minimax Method of Pattern Discovery and Probability Determination, A. D. Little, Inc., Cambridge, Mass., 1972.

14 R. Christensen, Proc. IEEE 1973 Conf. on Cybernetics and Society, IEEE Systems, Man and Cybernetics Society, 73 CHO 799-7-SMC, p. 321, 1973.
15 W. J. Dunn and S. Wold, J. Med. Chem., (1978) in press.
16 W. J. Dunn and S. Wold, J. Med. Chem., (1978), in press.
17 C. Mukherjee, M. C. Caron, D. Mulliken and R. J. Lefkowitz, Mol. Pharmacol., 12 (1976) 16.
18 R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York, 1977.
19 H. Wold, in R. Henn and O. Moeschlin (Eds.), Mathematical Economics and Games Theory, Springer, Berlin, 1977.

Analytica Chimica Acta, 103 (1978) 445—457
Computer Techniques and Optimization

THE USE OF AUTOCORRELATION TECHNIQUES FOR SELECTING OPTIMAL SAMPLING FREQUENCY APPLICATION TO SURVEILLANCE OF SURFACE WATER QUALITY

P. J. W. M. MÜSKENS

Catholic University of Nijmegen, Department of Analytical Chemistry, Faculty of Sciences, Toernooiveld, Nijmegen (The Netherlands)

(Received 3rd May 1978)

SUMMARY

The process monitoring system described is intended for use with processes where it is important not to exceed preset threshold levels. The autocorrelation function of the process, which is assumed to be a first-order stochastic process, is used to omit superfluous analyses. For processes with high autocorrelation values, many analyses can be omitted. The applicability of the method to analyses for the ammonium, nitrate and nitrite concentrations in the river Rhine over the period 1971—75 is reported.

Water quality surveillance requires not only information on the daily or hourly state of, for instance, a river; another goal of monitoring is to warn when the state of a surface-water system is likely to become critical. Such a situation might occur when the concentration of a certain component exceeds a maximum allowable or threshold value. The establishment of this threshold value is a problem which has to be solved by biochemists, biologists or environmentalists. The choice of the optimal measuring system for detecting whether or not the concentration has exceeded a fixed threshold is an analytical problem. The aim of the investigation reported here was to provide statistical methods that would be useful in selecting an optimal scheme for measurements. The most reliable scheme is achieved by measuring continuously, so that any values exceeding the threshold are detected immediately. However, a considerable number of these analyses may be superfluous, and savings in cost can be achieved by omitting all but the analyses which are really essential.

BASIC THEORY

The theoretical considerations required for decisions on how many analyses are superfluous are based on the assumption that the variation with time of the concentration of a given component in a river can be regarded as a first-order stochastic process [1]. The statistical properties of such a process are the mean value (μ_x), the standard deviation (σ_x), and the time constant (T_x).

The time constant T_x of a process can be estimated from the autocorrelation function, which is calculated from observed values for the process by means of the algorithm [1]

$$\phi_{xx}(\tau) = \left[\sum_{k=1}^{M-\tau} x(k)\, x(k+\tau)\right] \Big/ \left[\sigma_x^2\,(M-\tau-1)\right] \tag{1}$$

where $x(k)$ is the difference between the observed value at time k and the process mean μ_x. The value of $\phi_{xx}(\tau)$ quantifies the correlation between process values that are separated by a time lag τ. For $\tau = 0$, this correlation is perfect and $\phi_{xx}(0) = 1$. For increasing τ values the correlation decreases to $\phi_{xx}(\tau) = 0$. Many real situations can be represented as first-order linear processes which yield an exponential autocorrelation function [2]: $\phi_{xx}(\tau) = \exp(-\tau/T_x)$.

Proposed monitoring system

In the proposed system, a measurement is considered to be necessary if the information already available does not guarantee sufficiently that the process value will be lower than the threshold value. If the process value at time t is known, then the probability that the value at time $t + \mathrm{d}t$ will be less than the threshold value, can be quantified by using the conditional probability distribution function [3]

$$P[x(t+\mathrm{d}t) \mid x(t)] = \frac{1}{\sigma_p\,(\mathrm{d}t)(2\pi)^{\frac{1}{2}}} \exp\left[-\frac{\{x(t+\mathrm{d}t) - x_p\,(t+\mathrm{d}t)\}^2}{2\,\sigma_p^2(\mathrm{d}t)}\right] \tag{2}$$

where

$$x_p\,(t+\mathrm{d}t) = \exp(-\mathrm{d}t/T_x) \cdot x(t) \text{ and } \sigma_p(\mathrm{d}t) = \sigma_x\ \{1 - \exp(-2\mathrm{d}t/T_x)\}^{\frac{1}{2}} \tag{3}$$

(Symbols are listed in Table 1.)

The most probable value of $x(t + \mathrm{d}t)$ is given by $x_p(t + \mathrm{d}t)$, which can be called the predicted process value or prediction; $\sigma_p(\mathrm{d}t)$ is defined as the prediction error, and $\mathrm{d}t$ the prediction time. For increasing prediction times, the predicted value gradually approaches the mean process value, and the prediction error approaches the process standard deviation.

From eqn. (2), the P% reliability interval of the prediction is confined by the upper and lower limits

$$x_p(t+\mathrm{d}t) + N(P) \cdot \sigma_p(\mathrm{d}t) \text{ and } x_p(t+\mathrm{d}t) - N(P) \cdot \sigma_p(\mathrm{d}t) \tag{4}$$

where the reliability factor, $N(P)$, determines the probability, P, that the real process value $x(t + \mathrm{d}t)$ will be within the reliability interval. The probability that the process value will be outside the limits of the reliability interval, is given by $(1 - P)$. As σ_p increases with $\mathrm{d}t$, the reliability interval widens. This is used for the threshold monitoring system.

The operation of the system may be explained by a hypothetical example (Fig. 1). The process values $x(t)$ are assumed to fluctuate around the mean value μ_x, which is for convenience set at zero. The level G denotes the threshold value. The probability that a process value will be less than G is denoted by

TABLE 1

List of symbols

Symbol	Meaning
dt	Time unit, prediction time
$\phi_{xx}(\tau)$	Autocorrelation function of a process variable x
G	Threshold value
G'	Normalized threshold value = $(G - \mu_x)/\sigma_x$
m_a	Mean interanalysis time estimated from simulations
μ_a	Theoretically calculated mean interanalysis time
μ_x	Mean process value
$N(P)$	Reliability factor
P_g	Probability that a process value is less than G
P_u	Reliability of the monitoring system
P_u'	$(1 - P_u)$ = average probability of an unnoticed value exceeding the threshold
$P[T_a = idt]$	Probability distribution of the interanalysis time
s_a	Standard deviation of T_a obtained by simulation
σ_a	Standard deviation of T_a calculated theoretically
$\sigma_p(dt)$	Error of prediction
σ_x	Standard deviation of the process variable x
σ_z	Error of measurement
s_y	Observed standard deviation of real data series
T_a	Interanalysis time
T_x	Time constant of process variable x
τ	Time lag
$x(t)$	Process value at time t
$\bar{y}$	Observed mean concentration value

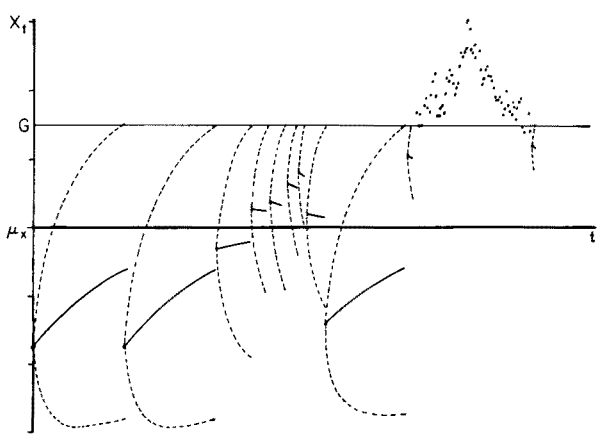


Fig. 1. Graphical representation of the operation of the monitoring system. The measurements are indicated by dots. (——) is the predicted process value or prediction. (– – –) indicates the reliability interval of the prediction. Here the 95% interval is used.

P_g, whereas $(1 - P_g)$ is the probability that a process value will exceed G. At time $t = 0$, the first measurement is made. From the result obtained, the reliability interval can be estimated as a function of time. As long as the upper limit of this interval does not reach level G, the probability of a result exceeding the threshold is less than the predetermined value $(1 - P)$, and a further measurement is not necessary. However, as soon as the upper limit exceeds level G, a measurement must be made, because the guarantee that the process value will not exceed the threshold is inadequate. From this new result, the time for the next measurement can be estimated similarly.

With this recursive procedure based on the most recent information, the measurement frequency will be lowered automatically if the process values are remote from the threshold. Conversely, the frequency will be increased when the process values approach level G. The timing of the measurements is fixed by the process properties, the threshold value and the reliability factor. As the process properties are inherent to the process and the threshold level is fixed by biological or environmental conditions, the only degree of freedom is the reliability factor $N(P)$. The higher the $N(P)$ factor, the lower the probability of undetected process values exceeding the threshold, but the higher the measuring frequency. The optimal value of the reliability factor can be determined by balancing the costs of the analyses against the costs caused by undetected values in excess of the threshold.

Determination of the frequency of analysis

In order to determine the frequency of analysis, the process is assumed to comprise M $(\to \infty)$ values, that are defined only at discrete times $k\mathrm{d}t$ $(k = 1, 2, \cdots, M)$. The interanalysis time and the time constant T_x are expressed in dt units; dt is the smallest possible value of the interanalysis time T_a.

An observed process value $x(t)$ will be followed by an interanalysis time T_a less than or equal to i dt, if the value $x(t)$ is larger than $q(i)$. This value can be determined by setting the upper limit in expression (4) equal to G and substituting eqns. (3)

$$q(i) = \mu_x + [G - \mu_x - N(P)\,\sigma_x\{1 - \exp(-2i\mathrm{d}t/T_x)\}^{\frac{1}{2}}]/\exp(-i\mathrm{d}t/T_x) \qquad (5)$$

For an interanalysis time of $i\mathrm{d}t$ the process value $x(t)$ must be within the interval $q(i)$—$q(i-1)$. The probability that this will happen is derived from the overall distribution function.

$$P[q(i) < x(t) \leqslant q(i-1)] = \frac{1}{\sigma_x (2\pi)^{\frac{1}{2}}} \int_{q(i)}^{q(i-1)} \exp[-\{x(t) - \sigma_x\}^2/2\,\sigma_x^2]\,\mathrm{d}x(t) \qquad (6)$$

The monitoring system will not include all process values in the interval betwee $q(i)$ and $q(i-1)$. The number of process values lying within this interval and actually used for measuring, can be computed approximately from

$$M\,P[q(i) < x(t) \leqslant q(i-1)]/i \qquad (7)$$

This formula can be made plausible as follows. The probability that a process value lies within the interval between $q(i)$ and $q(i-1)$ is given by eqn. (6). A process value in this interval will automatically be followed by an interanalysis time $i\,\mathrm{d}t$, thus a fraction $1/i$ of the process values in the specified interval will in fact be used for measuring.

The probability that the interanalysis time T_x equals $i\,\mathrm{d}t$ is then

$$P[T_a = i\,\mathrm{d}t] = P[q(i) < x(t) \leqslant q(i-1)]\,/i\left\{\sum_{j=1}^{\infty} P[q(j) < x(t) \leqslant q(j-1)]\,/j\right\}^{-1} \quad (8)$$

From the distribution function of T_a, the mean value (μ_a) and the standard deviation (σ_a) are computed.

$$\mu_a = \sum_{i=1}^{\infty} P[T_a = i\,\mathrm{d}t]\; i\,\mathrm{d}t \quad (9)$$

and

$$\sigma_a = \left\{\sum_{i=1}^{\infty} P[T_a = i\,\mathrm{d}t](i\,\mathrm{d}t)^2 - \mu_a^2\right\}^{\frac{1}{2}}$$

The mean frequency of analysis is defined as the reciprocal value of the mean time between two analyses. The quantities mentioned above must be determined numerically with a computer.

Figure 2(a) shows the frequency of analysis plotted as a function of the reliability factor $N(P)$ for four time constants and four threshold values; for convenience, a normalized threshold value G' defined as $(G-\mu_x)/\sigma_x$ is used. For $N(P) < G'$, measuring is unnecessary, because of the a priori probability $(1-P_g) < (1-P)$, which level must be attained before a new analysis is done.

The frequency of analysis increases with $N(P)$, the increase being faster for lower time constants. For a zero time constant, the frequency of analysis is either zero when $N(P) < G'$, or 1 when $N(P) > G'$. The time constant of the process is clearly an important property with regard to the monitoring system; the larger it is, the lower the frequency of analysis. For example, if $G = \mu_x + \sigma_x$, which implies that 16% of the process values exceed G, and $N(P) = 2$, and $T_x = 1$, the mean frequency of analysis will be 1. For $T_x = 10$, this frequency is reduced to 0.6, and for $T_x > 50$, the frequency of analysis is less than 0.4, which means that over 60% of the process values may be omitted.

Figure 2(b) shows the relationship between the mean frequency of analysis and the reliability factor $N(P)$ for the two-sided threshold problem, i.e. where there are two threshold levels symmetrical with respect to the process mean. Here the mean frequencies of analysis are much higher than for the corresponding one-sided problems.

Quantification of the reliability

In the monitoring system, $N(P)$ is used to fix the maximum allowable probability $(1-P)$ that the threshold may be exceeded without the need for

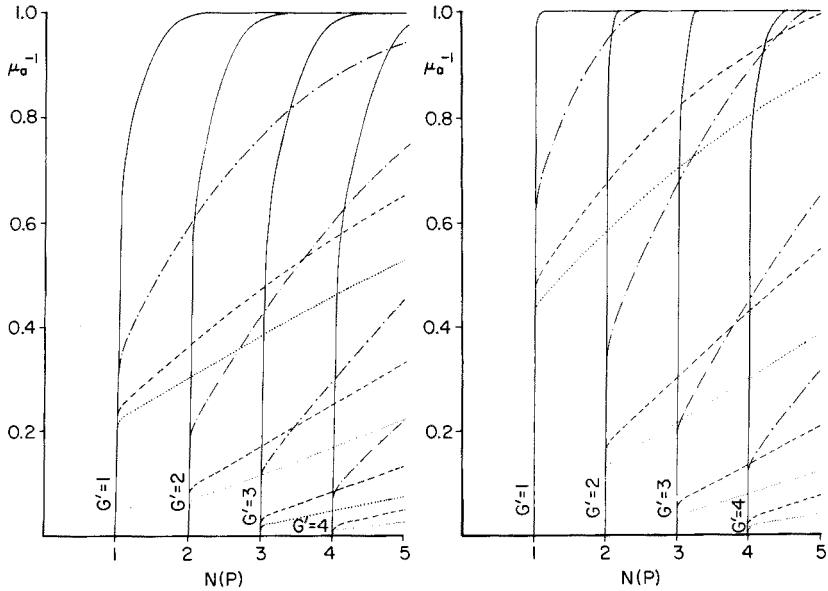


Fig. 2. The mean frequency of analysis ($1/\mu_a$) depicted as a function of the reliability factor $N(P)$, for four time constants T_x: 1 (——), 10 (·····), 50 (- - -) and 100 (.....), and four threshold values G': 1, 2, 3 and 4. Part (a) concerns the one-sided problem, i.e. there is only one threshold either above or below the mean value. Part (b) concerns the two-sided problem, i.e. the threshold levels are symmetrical with regard to the process mean.

an analysis. If the probability is higher than the preselected maximum, the guarantee that the process value will remain below the threshold limit will be inadequate. However, as long as the upper level of the reliability interval (4) does not reach the threshold value G, the probability of exceeding G is less than $(1 - P)$ and depends on the distribution function (2). Therefore this probability will vary with time, and an average probability must be calculated.

For each possible interanalysis time T_a, the average probability that the threshold will be exceeded is given by

$$Q(i\,\mathrm{d}t) = (i-1)^{-1} \prod_{j=1}^{i-1} \int_{x=-\infty}^{G} \exp\left[-\frac{\{x(t) - x'_p(j)\}^2}{2\,\sigma_p^2(j\,\mathrm{d}t)}\right] \mathrm{d}x(t) \qquad (10)$$

To prevent an underestimate of $Q(i\,\mathrm{d}t)$, $x'_p(j)$ is assumed to be predictable from the maximum value which can be followed by $T_a = i\,\mathrm{d}t$.

$$x'_p(j) = q(i-1)\exp(-j\,\mathrm{d}t/T_x) \qquad (11)$$

The average probability of the threshold being exceeded unnoticed, calculated over all possible interanalysis times, is:

$$P'_u = \sum_{i=1}^{\infty} Q(i\,\mathrm{d}t)\,P[T_a = i\,\mathrm{d}t] \qquad (12)$$

$P_u = (1 - P'_u)$ quantifies the probability that values in excess of the threshold will be detected. The higher the P_u value, the more reliable the monitoring system.

Figure 3(a) shows the relationship between the reliability P_u and the reliability factor $N(P)$ for the same time constants and threshold values as were used in plotting the curves in Fig. 2. For convenience, the ratio $(1 - P_u)/(1 - P_g)$ is plotted as a function of $N(P)$. This ratio indicates the fraction of the values in excess of the threshold that are undetected. If this ratio equals 1, values above the threshold are not detected; this will happen if no analyses are done. If the ratio equals 0, all values above the threshold are detected, which is only possible with continuous measurement. For $N(P) > G' + 1.5$, over 99% of the values above the threshold will be detected.

Figure 3(b) shows the ratio $(1 - P_u)/(1 - P_g)$ plotted against $N(P)$ for the two-sided problem.

Optimization

The reliability factor $N(P)$ is the only degree of freedom for the monitoring system (see above). Figure 3 shows that the reliability P_u increases with $N(P)$; consequently, the costs caused by undetected values over the threshold decrease. However, Fig. 2 shows that the frequency of analysis increases with $N(P)$, and therefore the costs of the analyses increase.

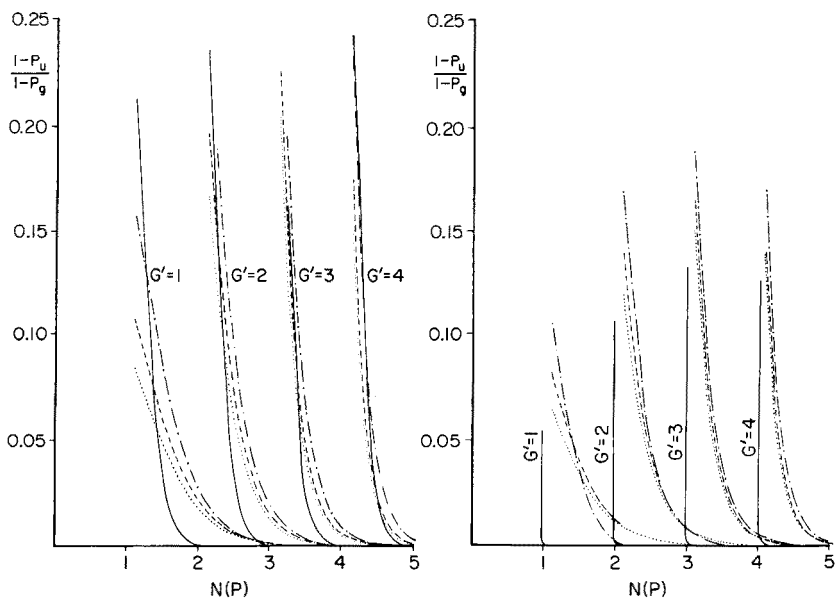


Fig. 3. The fraction of undetected values outside the threshold denoted by the ratio $(1 - P_u)/(1 - P_g)$ plotted as a function of the reliability factor $N(P)$, for four time constants T_x: 1 (——), 10 (·····), 50 (- - -) and 100 (.....), and four threshold values G': 1, 2, 3 and 4. Part (a) concerns the one-sided problem, and Part (b) the two-sided problem.

The optimal value of $N(P)$ is obtained for minimal total costs. If the costs of analyses and the costs caused by undetected values over the threshold are proportional to the mean frequency of analysis and the probability of undetected values over the threshold, respectively, a simple and elegant method can be used to find the minimal total costs. For this purpose, the reliability of the monitoring system is plotted as a function of the frequency of analysis (Fig. 4). The optimum frequency of analysis and reliability is found directly from the coordinates, where the curve has a slope of $-K_a/K_b(1-P_g)$; K_a represents the costs per analysis and K_b the costs per undetected value above the threshold. For zero K_a, the optimum frequency is 1, i.e. continuous measurement. For zero K_b, the optimum frequency is zero, i.e. analysis is unnecessary.

In Fig. 4(b), the relationships between the ratio $(1-P_u)/(1-P_g)$ and the mean frequency of analysis calculated for $T_x = 1$ are not shown, because they comprise only two points with the coordinates (1.0, 0.0) and (0.0, 1.0). For $N(P) < G'$, the mean frequency of analysis is zero and the ratio $(1-P_u)/(1-P_g)$ is one; for $N(P) > G'$ the frequency is close to one whereas the ratio $(1-P_u)/(1-P_g)$ is approximately zero.

If part of the total costs does not fulfil the conditions of linearity mentioned above, a more elaborate method of establishing the optimum is necessary. Simplex optimization is useful in this case [4].

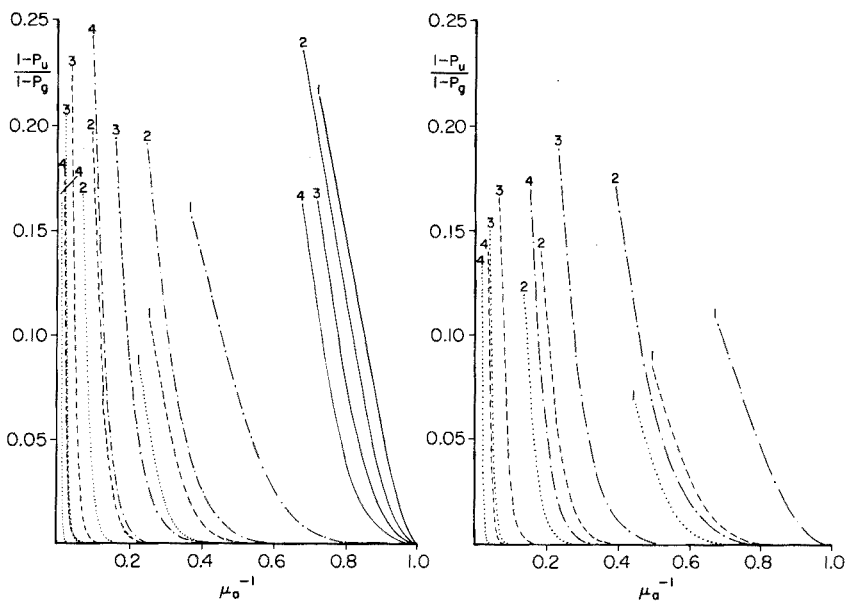


Fig. 4. The fraction of undetected values outside the threshold denoted by the ratio $(1-P_u)/(1-P_g)$ plotted as a function of the mean frequency of analysis $(1/\mu_a)$, for the four time constants and threshold values indicated in Fig. 3. Parts (a) and (b) relate to the one-sided and two-sided problems, respectively.

RESULTS AND DISCUSSION

In the above theoretical calculations of the distribution of the interanalysis time, some approximations were required. Their validity was examined by calculating the mean value and standard deviation of the interanalysis time and the system reliability for some values of T_x, G' and $N(P)$. The results were compared with the corresponding results obtained by simulation. The practical applicability of the system itself and the theory were tested by applying the monitoring system to real data obtained from observations of the concentration of ammonium, nitrate and nitrite ions in the river Rhine.

Test of the theory with simulations

For testing the above theory, discrete processes were simulated by means of a discrete white noise generator [5]. The time constants T_x and standard deviations σ_x were selected by the procedure of Gelb and Palosky [6]. Processes with a zero mean and unit standard deviation were simulated for three time constants, T_x = 10, 50 and 100. For each process 10 000 values were simulated. The monitoring system was applied to the processes simulated for threshold values, G, of 1, 1.5, 2, 2.5 and 3 and reliability factors, $N(P)$, of 1, 1.5, 2, 2.5, 3 and 4, with the restriction that $N(P) > G$.

In Fig. 5(a), the values $1/\mu_a$ calculated from expression (9) are compared with the simulated values $1/m_a$. The mean value m_a and standard deviation s_a were determined from the simulations by

$$m_a = \frac{1}{n}\sum_{l=1}^{n} T_a(l) \text{ and } s_a^2 = \frac{1}{n-1}\sum_{l=1}^{n}\{T_a(l) - m_a\}^2 \tag{13}$$

where n is the number of measurements and $T_a(l)$ is the interanalysis time after the lth measurement. The theoretical values are seen to be systematically lower than the simulated values. This is a direct consequence of the approximation in expression (7), which appears to underestimate the mean interanalysis time. Furthermore, the standard deviation σ_a increases with increasing μ_a, which also causes deviations between the calculated and observed mean values of μ_a. However, for practical applications, it can be concluded that the theoretical mean frequency of analysis is in good agreement with the mean frequency observed empirically. Figure 5(b) shows the comparison between theory and simulation for the two-sided problem. In order to test the formulae used in calculation of the reliability, the values of $(1 - P_u)$ computed from eqn. (12) were compared with the corresponding values obtained by application of the monitoring system to the simulated processes (Fig. 6a). Deviations from the ideal relationship are considerable. However, as the number of values exceeding the threshold is only a small fraction of the simulated process values, the fraction obtained by simulation could vary by a factor of 5 from the fraction calculated theoretically. Better results would have been obtained by simulating processes over a longer period, e.g. some multiple of the 10 000 values used here. However, it can be concluded that the theoretical and

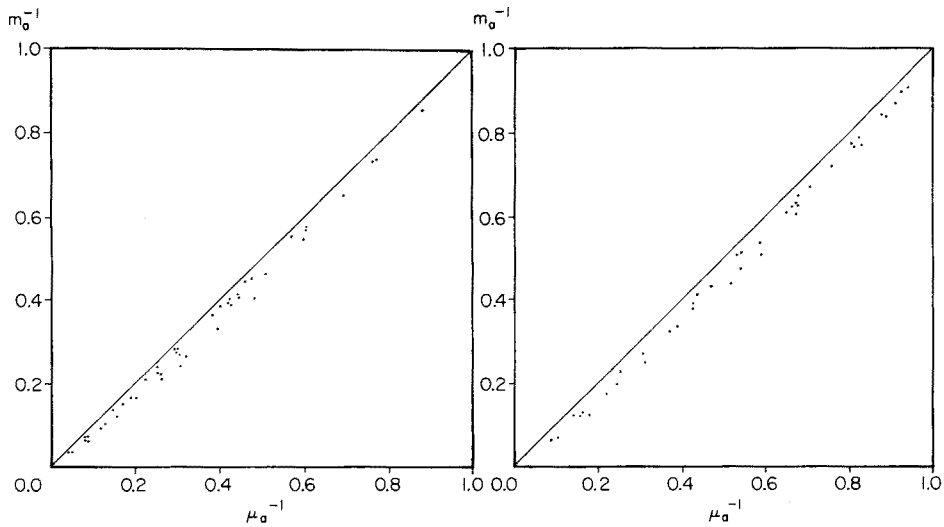


Fig. 5. Comparison of the mean frequency of analysis computed theoretically ($1/\mu_a$) with the value estimated from simulations ($1/m_a$). Parts (a) and (b) relate to the one-sided and two-sided problems, respectively.

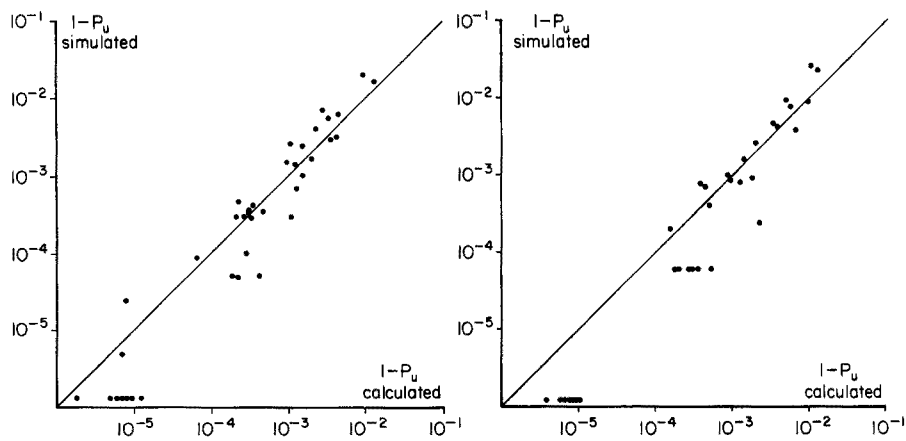


Fig. 6. Comparison of the probability of an undetected value above the threshold computed theoretically (horizontal axis) with the corresponding value estimated from simulations (vertical axis). Parts (a) and (b) relate to the one-sided and two-sided problems, respectively.

simulated values agree sufficiently for practical applications. Figure 6(b) is drawn for the two-sided problem.

Application to practical situations

A problem that is very suitable for practical application of the proposed monitoring system is the surveillance of the quality of surface waters. This

investigation was confined to surveillance of the concentrations of ammonium, nitrate and nitrite ions, which are measured in the river Rhine at the Bimmen Water Control Station. [The data used were provided by the Landesanstalt für Wasser und Abfall, Düsseldorf, Nordrhein-Westfalen (B.R.D.).]

The statistical properties of the time-dependent variations in the concentrations were estimated from daily measurements during the period 1971—1975 (Table 2). As was shown earlier [7], the observed values y can be regarded as superpositions of the process values x and errors of analysis z. Thus, the total standard deviation s_y is a combination of the process standard deviation s_x and the standard deviation s_z of the analytical method. T_x is the estimated time constant of the process.

The skewness and kurtosis [8] are properties which depend on the probability distribution function, and are, respectively, 0 and 3 for normal distribution. In practice, deviations from the ideal values always occur, and therefore must be tested for significant difference from the ideal values [9]. However, these tests do not take autocorrelation into account, so that only qualitative statements can be made. The results (Table 2) indicate that the concentrations are not normally distributed. Consequently, the monitoring system, derived for normal distributions, cannot be used in a straightforward way, and the data have to be transformed. In this investigation, all variables are transformed to their logarithms. The statistical properties of the logarithmic concentrations are given in Table 2. The mean value and standard deviations are altered considerably, but the process time constant T_x is hardly affected by taking logarithms The skewnesses of the logarithms approach the values required for normality more than those of the original observed values. The kurtoses of the logarithms for nitrate and nitrite are still large. However, as kurtoses larger than 3 imply that the distribution is sharper than normal, the reliability of the monitoring system is greater than for a normal distribution. For kurtoses less than 3, the distribution would be flatter, and the threshold would be exceeded more often than normal distributions allow. These considerations indicate that the logarithmic values are preferable to the original values for application in the monitoring system.

The monitoring system was applied to the logarithms for a normalized threshold value $G' = 2$, with reliability factors $N(P) = 3$ and 4, and $G' = 3$, with $N(P) = 4$. Both one-sided and two-sided threshold problems were investigated. With application to real data, measurement errors must be taken into account. The process value $x(t + \mathrm{d}t)$ can be predicted from the measurement value $y(t)$ by means of $x_p(t + \mathrm{d}t) = \exp(-\mathrm{d}t/T_x) \cdot y(t)$. The prediction error can be calculated to be:

$$\sigma_p(\mathrm{d}t) = \sigma_x \left[1 - (1 - \sigma_z^2/\sigma_x^2) \exp(-2\mathrm{d}t/T_x)\right]^{\frac{1}{2}}$$

and substitution of this equation in the appropriate formulae of the Theory section. The theoretical values of μ_a and σ_a are compared with the empirical values obtained by application of the monitoring system to the logarithms of the concentrations of ammonium, nitrate and nitrite ions in Table 3. It is clear

TABLE 2

Statistical properties of the observed and logarithmic concentrations of ammonium, nitrate and nitrite ions in the river Rhine
(For a normal distribution the skewness is 0 and the kurtosis is 3.)

Property	Observed			Logarithmic		
	NH_4^+	NO_3^-	NO_2^-	NH_4^+	NO_3^-	NO_2^-
Mean value, $\bar{y}$ (mg l^{-1})	3.07	13.43	0.43	0.93	2.53	−0.92
S.d., s_y (mg l^{-1})	1.97	5.02	0.16	0.65	0.36	0.39
T_x (days)	100	90	29	110	70	31
s_x (mg l^{-1})	1.87	4.65	0.15	0.62	0.34	0.36
s_z (mg l^{-1})	0.62	1.88	0.05	0.18	0.14	0.14
Skewness	1.56	1.58	0.75	−0.34	−0.44	−0.69
Kurtosis	5.98	8.16	4.05	3.22	5.62	5.74

TABLE 3

Mean values and standard deviations of the interanalysis time (expressed in days) determined theoretically and empirically for the threshold monitoring of the (logarithmic) concentrations of ammonium, nitrate and nitrite ions

G'	$N(P)$	μ_a	m_a	σ_a	s_a
Ammonium ion					
+2	3	4.80	5.51	11.95	12.56
−2	3	4.80	5.60	11.95	12.53
+2 & −2	3	2.52	2.93	4.30	4.65
+2	4	3.12	2.87	6.29	5.59
−2	4	3.12	3.11	6.29	6.21
+2 & −2	4	1.67	1.49	1.83	1.29
+3	4	11.04	10.80	19.31	19.78
−3	4	11.04	9.26	19.31	18.68
+3 & −3	4	6.04	5.32	9.19	8.57
Nitrate					
+2	3	3.05	4.33	7.21	8.69
−2	3	3.05	3.32	7.21	7.66
+2 & −2	3	1.61	1.71	1.99	1.86
+2	4	2.00	1.89	3.57	3.17
−2	4	2.00	1.77	3.57	3.21
+2 & −2	4	1.11	1.00	0.42	0.00
+3	4	5.31	8.95	10.45	13.78
−3	4	5.31	8.26	10.45	12.87
+3 & −3	4	2.85	4.01	4.26	4.76
Nitrite					
+2	3	2.72	2.91	4.30	4.61
−2	3	2.72	2.97	4.30	4.42
+2 & −2	3	1.51	1.46	1.23	1.09
+2	4	1.82	1.63	2.13	1.92
−2	4	1.82	1.70	2.13	1.86
+2 & −2	4	1.08	1.00	0.28	0.00
+3	4	4.33	4.75	5.82	5.86
+3	4	4.33	4.56	5.82	5.61
+3 & −3	4	2.47	2.60	2.45	2.33

that the empirical results agree very well with the corresponding theoretical values. The deviations, e.g. for nitrate, are caused by two factors: (1) the time from which m_a and s_a are estimated is too short; (2) some outlying values caused by measurement errors occur in the observed data.

Conclusions

When the proposed monitoring system is applied, analytical measurements are unnecessary when the process values are likely to be far from the threshold levels, and the frequency of analysis is low. When the process values approach the threshold level, the frequency of analysis is increased. The monitoring system is based on the autocorrelation function, which is used as a prediction function. The mean frequency of analysis and the reliability of the monitoring system can be calculated theoretically, so that an optimal reliability factor can be selected. The theoretically calculated frequency of analysis and reliability agreed well with empirical values obtained from simulated as well as real processes.

The introduction of this system in practical situations, e.g. in surveillance of surface water systems, should lead to cost savings in the number of analyses required.

REFERENCES

1 G. E. P. Box and G. M. Jenkins, Time Series Analysis, Holden-Day, San Francisco, 1970.
2 F. A. Leemans, Anal. Chem., 43 (11) (1971) 36A.
3 I. F. Blake and W. C. Lindsey, IEEE Trans. Inform. Theory, 19 (1973) 295.
4 S. N. Deming and S. L. Morgan, Anal. Chem., 45 (1973) 278A.
5 J. A. G. M. Kerbosch and R. W. Sierenberg, Discrete Simulatie, Samsom, Alfen aan de Rijn, 1973.
6 A. Gelb and P. Palosky, IEEE Trans. Aut. Control, 11 (1966) 148.
7 P. J. W. M. Muskens and W. G. J. Hensgens, Water Res., 11 (1977) 509.
8 O. L. Davies and P. L. Goldsmith, Statistical Methods in Research and Production, Oliver and Boyd, Edinburgh, 1972.
9 R. D'Agostinho and E. S. Pearson, Biometrika, 60 (1973) 613.

Analytica Chimica Acta, 103 (1978) 459—468
Computer Techniques and Optimization

SMALL LABORATORY COMPUTER NETWORKS

RAYMOND E. DESSY
Chemistry Department, Virginia Polytechnic Institute, Blacksburg, VA 24061 (U.S.A.)
(Received 10th May 1978)

SUMMARY

The rational approach to laboratory automation in a changing research environment involves a network of computers. This permits program preparation and data manipulation/storage/plotting/printing to take place on the CPU most appropriate to the task. It also allows the application programs to run on satellite processors with minimum configuration and maximum data throughput capabilities. The hardware implementation is described, as well as a brief introduction to the software facilities available under RT-11/REMOTE and FORTH, the two language systems used at Virginia Polytechnic Institute to implement the net.

The art of analytical chemistry is being altered by the impact of computer technology. At one extreme, large data bases are being correlated to abstract more information from existing analyses, and these results are being reported in a format more adaptable to human senses by means of sophisticated graphics, software and hardware. At the other extreme, tasks involving large numbers of repetitive operations or in which large amounts of data are made available by each experiment have been successfully automated. The implication of this last word is NOT just that some computing element has been added to the system, but that the manual operations involved in instrument setup, calibration, sampling, and data collection have also been relegated to an electromechanical servant. Laboratory automation involves much more than just acquiring a computer and also more than collecting data from the instrument automatically.

In many laboratories, automation has been nucleated by a minicomputer serving several instruments, thus creating small islands of computer capability. These have grown to a critical and optimum size — and then have frozen in a fixed configuration. Financial and real-time constraints forbid an extension of their capabilities or resources. These systems continue to consume maintenance funds until amortization permits replacement.

Chemists frustrated by the lack of response to their automation needs are seeking redress in intelligent instruments. This will improve sample throughput, thus accommodating the current pressure on most laboratories. But record keeping and report writing will increase. However, this increased data flow demands that the extremes of computing technology, such as data base

management and input/output (I/O) control, be brought together so that they can interact.

A route to open-ended design for laboratory automation is evolving to permit computing facilities to share resources, capabilities, and data. Such a scheme allows small systems, with just sufficient hardware to accomplish a specific task, to be located near the experiments or chemical process. Yet this small system can call on and use all of the powerful facilities associated with larger computer systems during program development and debugging phases, including the mass storage elements needed for retrieval and correlation of large data bases accumulated via multi-instrument analytical services. Routine reports can be computer-generated.

This route is now becoming available in what is termed networking or distributed processing.

Viewed as separate entities, the three classes of automation tools — midi, mini and micro — are remarkably complementary. The larger midicomputers are systems that operate under the control of exceedingly complex software systems (operating systems). They have excellent computing power and can support vast amounts of memory. Expensive printers and plotters can be added to communicate effectively with the users who can justify their existence. These users may be laboratory managers who need access to distilled progress reports on the nature, amount, and quality of the data that their analytical groups are producing, or supervisors who need exposure to the financial and efficiency factors so important to proper system growth. Minicomputers operate under the control of much less sophisticated operating systems. Operating with minimum overhead in executing software, they can respond rapidly to the needs of their instruments; operating without large numbers of attending systems engineers and analysts, they can readily respond to the needs of their user. Microcomputers are the offspring of the large-scale integrated (LSI) circuit technology that produced hand-held calculators. These inexpensive devices can control motors, heaters, or valves; they can take data from push buttons, thumb-wheel switches, or joy-sticks; and can display data in lights, panel displays, or on TV tubes. The calculations required are minimal, and the need to expand is non-existent. These features have made them the designer's choice in creating intelligent instruments.

In many laboratories where separate development of automation schemes involving two or three of the above routes is on-going, panic is developing because of the limited or non-existent intercommunication between isolated systems. But in the several laboratories where responsible personnel are technologically competent and mutually cooperative, the synergic interaction of micro, mini, and midicomputer facilities has led to the development of an intercommunicating network of automated facilities. Classic examples of successful networking can be seen in industrial process control facilities such as Union Carbide or Phillip Morris, and in academic laboratories such as the University of Oregon, and the University of Hamburg. However, the inaccessibility of software or non-convertibility of software has precluded widespread dissemination of the techniques.

Major changes in the marketing strategy of the large computer vendors are rapidly altering this state of affairs. They have analyzed the computer markets that are rapidly saturating or becoming excessively competitive, and have developed an increased awareness of end-user needs. As a result, these vendors are beginning to provide almost turn-key network hardware/software packages in a variety of forms.

The question of WHY?, WHAT?, and HOW? will be explored so that one can judge his own laboratory's needs for networking capabilities. This is done, not with the expectation that the average chemist will implement the net, but rather with the firm conviction that the laboratory personnel must be the most important factor in determining the final state of laboratory automation. To do this requires that a vocabulary and philosophy be acquired so that the reader can have an affective and effective voice in the decision.

The establishment of a network facility is indicated under a number of etiologies:

(a) real, not imagined, fear of total dependence on one computer;

(b) acquisition of a number of intelligent instruments which leads to increased throughput which, in turn, leads to a data correlation and report generation bottleneck;

(c) addition of another instrument to a multi-instrument minicomputer-based system which taxes the real-time capability of the central processing unit (CPU);

(d) a minicomputer system becomes so burdened with the need for real-time data that the background tasks of data correlation and report generation become backlogged, the system response degrades, and software development costs escalate;

(e) a midicomputer system that has been successful in its role of a business-oriented machine is asked to handle the control and data acquisition needs of laboratory equipment.

By adding networked micro, mini, or midicomputer capabilities, each of these scenarios can be successfully resolved. Each of these examples could also be resolved by a non-network addition of computing/automation hardware. However, networking is the only route that will lead to an open-ended, expansible automation scheme. In it, the individual elements can be articulated, developing the full capabilities of each member, rather than forcing one member to duplicate inadequately the ability of another.

The elements of growth potential and intercommunication between elements are the heart of a computer network. At this point it is appropriate to differentiate between network software/hardware and the corresponding elements in distributed processing. The model is that of Dr. Richard Eckhouse, Digital Equipment Corporation. The concepts and philosophies of the interrelations, interactions and cooperation between processors can be labelled "coupling". This coupling may be "loose" or "tight" depending on the design and function of the system. Four categories can be cited:

	Hardware logic interconnections	Software logic interrelations
Multiprocessors	Tight	Tight
Networks	Loose	Loose
Distributed processing	Loose	Tight
Most aggregations of computers	Fallible	Incoherent

In multiprocessor environments the two (or more) CPU's run under one operating system, and are constantly aware of the state of each other. In networks, programs and data are shipped from one site to another, and resources are shared, but each processor remains a highly self-centered node, much like individual cells in a tissue. They can sense changes and needs in the surrounding cells, and can intercommunicate with them; but there is no global awareness. This comes with distributed processing. Software for such implementations is under development. For this reason, the rest of this article focuses on networks that are useful in the science laboratory.

In a microcomputer or intelligent instrument, a stored program determines what actions are to be taken. The stored program has only three functions: to interact, if necessary, with the operator of the instrument so as to perform the analysis correctly; to perform the data acquisition operation; and to provide a usable report to the user. The "smart" g.c. and l.c. equipment currently being vended commonly performs peak picking, sample identification, and quantification based on user input and vendor algorithms. Some vendors also provide the capability of transmitting either the entire or the distilled data file from the intelligent instrument to a more sophisticated computer. This will become increasingly valuable as networks develop.

It is quite common to have intelligent instruments coexist in a laboratory equipped with special analytical devices automated by small computers. However, there is no simple way to correlate the output of these various instrument other than having a chemist or technician monotonously transcribe the data into a compiled list.

It is possible to interconnect computer elements, so that laboratory computers can transmit their data to a larger host computer. It would then be a trivial matter for the host computer, with sufficient rotating memory, to store the output from each instrument over long periods of time, to implement searches through each list of data, and to pair up matching sample numbers.

In many laboratories and particularly in process control environments, the intelligent equipment near the experiment often becomes insufficient for the task. Under changing conditions, different decision processes must be used, which require different programs. This process can be augmented by programs running in the host which make decisions about the operation of the satellite and automatically change its programs.

As experiments become more sophisticated and the time and function demands made on computerized systems increase, many analytical chemists will want more than an intelligent instrument on their benches. They would

like to develop special function programs and use the facilities of a larger system for printing, plotting, and manipulating their data.

In the smaller computers, software development costs represent over 80% of the installation cost. As hardware decreases in price and software increases, it becomes important that software development aids be used extensively. These tools are available on mini and midicomputers running good operating systems, and support the CRT terminals, line printers, etc., which speed programming and improve documentation standards. It is possible to create programs on these host systems in Assembler, BASIC, or FORTRAN in such a way that they can be loaded down-line in machine code to a satellite and executed. In the satellite, they may be executed and debugged without fear of bringing the main system down; yet, when fully operational, the satellite can use all the resources of the machine that precedes it. This brings us to the point of trying to understand how a network operates for the chemist.

Our laboratory is currently involved in research dealing with the design and testing of automated instruments. This involves preparing programs for the new instruments, and then monitoring their activities over extended time periods to evaluate the nature and quality of the data they produce. The network facilities to accomplish this task consist of interconnected host processors, and eight satellite computers (Fig. 1). The two host computers are involved with program preparation, data correlation, and data display. The satellites are involved with data collection. The configuration is as outlined in Table 1. Block transfer between hosts allows facile data and program exchange.

Programs are prepared on the host computers in a mixture of appropriate languages. Assembler, FORTRAN, BASIC and FORTH are supported. The load modules are either stand-alone in character, or are linked with a simulated operating system module that can share the resources of the host computers as the programs are executed on the satellites. As programs are needed in the remote processors they are loaded down-line over RS-232 twisted pair lines at 9600 Baud. Although programs are normally debugged at the satellite, it is often necessary, when mixtures of high-level languages and Assembler are

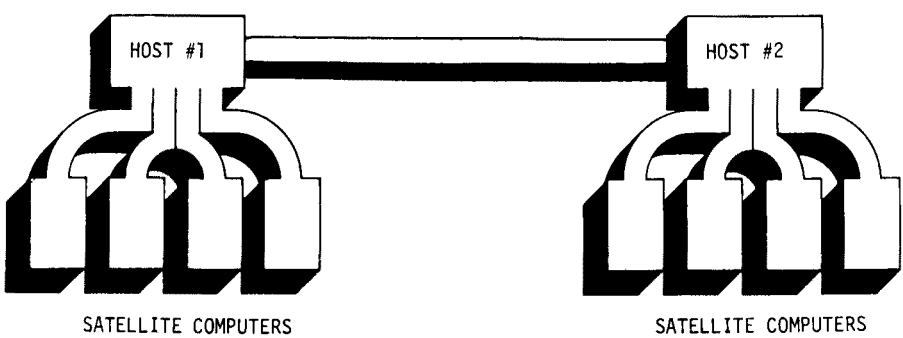


Fig. 1. Network facility consisting of interconnected host processors and eight satellite computers.

employed, to return to the host where extensive high-level debugging aids are available. The first characteristic of the system is thus development of program on adequately supported host processors, then down-line loading and debuggin in nodes where system crashes may be avoided.

Programs in the satellite computer acquire and massage the data, presenting, if necessary, local reports in terms of CRT displays, TTY output, or X/Y graphic plots on CRT or video terminals. With vendor supplied FFT programs (in a laboratory applications package), exploratory Fourier transform electrochemical experiments have been implemented in a day, doing program preparation on the host where extensive libraries can be supported, followed by down-line loading and execution in the satellite where real-time data acquisitio at 30 kHz and forward and reverse FFT are supported. The second characteristic of a network is that real-time tasks with high speed requirements are run on processors with minimum software overhead making program generation and debugging simple and fast.

As data are collected by the various processors, they can be transmitted bac to the host for storage on disk. These transmissions can be elicited by calls similar to those that would be employed at the host if it were running the applications program. The third characteristic of this architecture is the sharing of resources. Once the files are on the host's rotating storage, it is possible for plotting routines to be applied, or record management routines can provide correlated sets of information to the chemist. Typical applications in this laboratory involve generating response—concentration, or response—time curves for the evaluation of transducers.

Many of the efforts in this laboratory require the development of software which will run on independent ROM-based processors. All of the CPU's are identical in the network. Thus programs can be developed on the host, loaded down-line, tested and debugged in the net, and then the acceptable program burnt into PROM for field use. Where the network uses processors with a common machine language, a final characteristic of the net is realized: transportability of machine language code.

Finally, with most of the major components of the system duplicated at various nodes, and in the hosts, the mean time needed to repair a malfunctioning system is low. Board swapping localizes the problem quickly, and movement of boards from a low demand site to a high demand site provides continuous operation.

The network system is currently running under RT-11/VO3, with the net implemented by REMOTE, which provides a sub-set of Decnet. A network totally implemented by FORTH has been generated. FORTH is a structured programming language that allows the user to develop his own vocabulary as the program is written. This approach, so different from the rigid structures of FORTRAN or BASIC, promises to lead to the development of languages far more suited to the laboratory environment than the classical high-level languages. In rapidly changing research environments, Assembler programming provides the flexibility needed, but at a programming speed and cost that is

prohibitive. Classical compilers and interpreters permit rapid program development, but become rigid as they attempt to provide real-time capabilities, variable file-formatting protocols and other "extensions" using as a base a language inappropriate to this task. Benchmarks in this laboratory indicate that FORTH will permit accessing of disks 40 times faster than FORTRAN in real-time data acquisition storage (DMA ADC, assembler acquisition program linked to FORTRAN program). Memory compressions of 20 are typical, and math execution speeds appear to be about a factor of 2 faster, because of the faster linkages and argument passages in FORTH. As a programming *system*, rather than just a *language*, FORTH does not suffer from the convolutions of mating which occur between an operating system and its higher-level languages.

Adding record management and data-base management facilities to such a net is not technically difficult. It is limited only by the lack of a clear analysis of the proper data-base management to use for each laboratory environment — and then by the cost of implementing it. It is this expansibility and flexibility of networks which makes it attractive as a rational plan for laboratory automation. Then, when true distributive processing is available and needed, the hardware will be already in place.

The heart of a computer network is the protocol conventions that it employs. These are the accepted conventions used to transmit information electrically and logically from one computer to another. These protocols are layered, one on another, to span the range from the hardware plugs used to interconnect units to the software commands that implement network functions. Most chemists will see only the outer layer of the onion, i.e. the software commands, but the entire core must be there. For example:

TABLE 1

Components of the network facility

	Host 1	Host 2	Typical satellite
CPU	LSI-11	LSI-11	LSI-11
MEMORY	28K Static Ram	28K Static Ram	12K mixed RAM and Core
DISK	Dual Fixed/Removable Moving Head (10 MByte)		
	Tri-Double Density Flexible Disks (2 MByte)	Tri-Double Density Flexible Disks (2 MByte)	
ADC's	16 Channel	16 Channel	16 Channel
DAC's	2 Channel	2 Channel	2 Channel
PARALLEL I/O	16 Bits In/Out	16 Bits In/Out	16-32 Bits In/Out
SERIAL I/O	6 Serial Ports	6 Serial Ports	2 Serial Ports
Plotting	100 cm Digital Plotter	20 x 30 cm Analog Plotter CRT display	CRT
Terminals	Video + DecWriter	Video + DecWriter	Video or TTY
Line Printer	300 lpm Mohawk	Dual 9 Track	
Tape Drives	800 bpi, 25 ips		

User command protocol — formats for user requests at a terminal.
Machine command protocol — formats of these requests in machine code.
Logical link protocol — formats that assure sequential reconstruction of individual messages.
Physical link protocol — formats that assure error-free transmission.
Hardware protocol — electrical formats used on the hardware interconnection.

In some networks all of the members are attached to a common party line or multi-drop (Fig. 2a). The destination responds only when a transmitted signal activates it, just like a party-line telephone, and other units disregard the signal. The IEEE Laboratory Bus works this way. Other networks have point-to-point connections between members, often in a star or hierarchical arrangement (Fig. 2b). This is akin to a private leased wire. In very large systems, multi-nodal arrangements occur with many possible pathways from source to destination (Fig. 2c). It is the function of the original handshaking message to find the best available route for the communication by trying various pathways and discarding them one-by-one if they are busy. Such networks can survive failures of one of the computer members with minimal loss of effectiveness.

The language used in the protocols involves the transmission of both data and control characters. In the IEEE Laboratory Bus, the control language has been created specifically for the system. Other vendors will use American Standard Code for Information Interchange (ASCII) or Extended Binary Coded Decimal (EBCDIC). The networks should be capable of handling all possible characters in these sets — 8 bits will permit 256 combinations. This insures that straight binary information can be transmitted.

At the hardware protocol level, the electrical form in which the data are transmitted is significant. Common methods of transmitting data involve parallel or serial routes. In the former, the bits representing the data word are shipped simultaneously down multiple-conductor cable. The IEEE Laboratory Bus is 16 bits parallel, 8 bits of control information, and 8 bits of data.

Serial transmission links involve sending one bit at a time of each byte down a twisted pair of lines. Asynchronous serial linkages are used in many

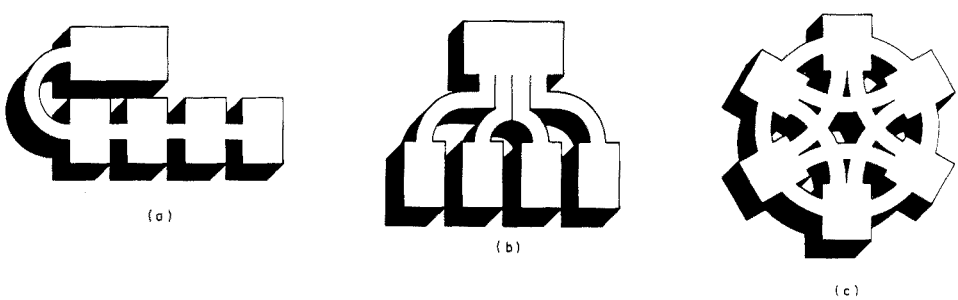


Fig. 2. (a) Multi-drop network; (b) point-to-point network; (c) multi-nodal network.

networks and are always found in intelligent instruments which offer computer compatibility.

Most intelligent instruments meet hardware protocols, but lack the proper logical protocols. It is often necessary to inject a microcomputer between them and a network.

The flexibility and expansibility of networking can be illustrated by a possible development pattern in a quality control — process control —research analytical laboratory in an environment common to chemical companies (Fig. 3).

In a polymer physical testing laboratory faced with tensile strength, elongation, cold flow, and molecular weight determinations, the IEEE Laboratory Bus has some interesting advantages because scanning analog-to-digital conversion (ADC) equipment, frequency meters, voltmeters, relay actuators, and printers are available from numerous vendors with the IEEE Laboratory Bus adapters built in. The microcomputer easily handles the control/data flow and report generation in the system, the programming being done on a host used for laboratory management purposes (Fig. 3a).

In a quality control laboratory, which analyzes samples from customers and from process control streams, intelligent instruments with serial ASCII

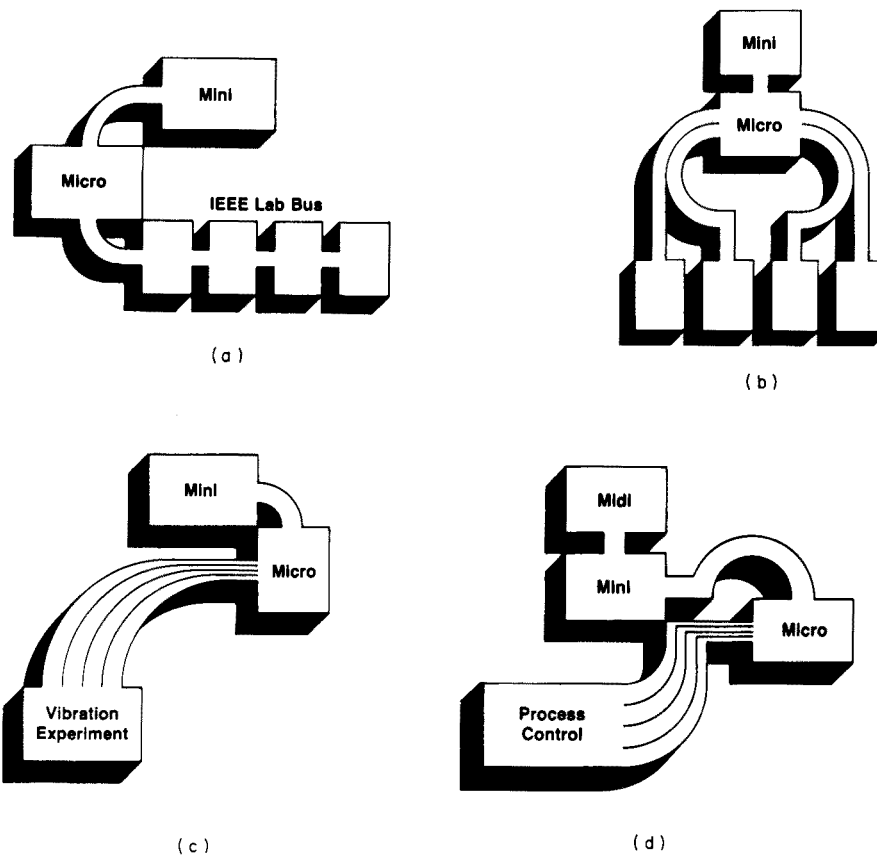


Fig. 3. Networks designed for: (a) polymer laboratory; (b) intelligent instruments quality control laboratory; (c) research laboratory; (d) process control environment.

RS-232/C outputs are attached to a microcomputer that allocates sample numbers and other identifying information, concentrates the data, and passes this to a larger host (Fig. 3b). In a research laboratory, data collection from a special vibration experiment is controlled by a microcomputer subsystem connected to a host computer where the actual Fourier transform is made (Fig. 3c).

Finally, in a process control environment (Fig. 3d), with over 500 sampling points for temperature, pressure and flow rate, and 200 stations for chemical analyses, a very high-speed (1 megabit s^{-1}) coaxial line connects a minicomputer to several microprocessor-based industrial data acquisition stations to provide local data-base management facilities. A midicomputer serves as a host to the minicomputer and as a satellite to the Corporate computer. With little effort and an open mind, the entire resources of a complex network will shortly be used as casually as a hand-held calculator is used today.

The research group at VPI & SU uses a network of two intercommunicating DEC PDP-11 hosts attached to four other satellite computers. This computer network has been developed by the author and M. Starling, W. Nunn, D. Hooley, H. Wohltjen, C. Knipe, D. Binkley, J. Berquist, G. Giss, E. Fiorino, J. Fiorino, C. Baker, I. Starling, and I. Bowater. The drawings are used by the courtesy and permission of *Analytical Chemistry*. Unrestricted grants from the Gillette Charitable and Educational Foundation have supported this development.

Analytica Chimica Acta, 103 (1978) 469—476
Computer Techniques and Optimization

CHEMICAL PROCESS OPTIMIZATION BY COMPUTER — A SELF-DIRECTED CHEMICAL SYNTHESIS SYSTEM

H. WINICOV,* J. SCHAINBAUM, J. BUCKLEY, G. LONGINO, J. HILL and C. E. BERKOFF

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pa. (U.S.A.)

(Received 3rd May 1978)

SUMMARY

A closed-loop automated chemical synthesis system has been designed for the purpose of optimizing chemical reaction parameters. This system uses a time-sharing computer with the simplex algorithm for optimization programmed in extended BASIC language. Strings of ASCII characters from the computer are trapped by terminal hardware at the reaction and analysis sites and are used to control all phases of the system.

Process development is a vital part of marketing any chemical substance. In some industries, such as the pharmaceutical industry, the turnover rate of new chemical syntheses is very high because of the large number of compounds being evaluated at any one time and the high attrition rate during the initial stages of testing. At the same time, it is important that the optimum process be identified at a relatively early stage in the development of a drug not only to obtain a competitive edge, but also to introduce into the testing cycle chemicals made by the "final" process. Thus, several processes must be optimized for each compound under study so that an informed decision can be made to select the best overall route with regard to starting materials and synthetic idiosyncrasies. Traditionally, this decision is made in part by intuitive processes (i.e. "experience") and in part by carefully controlled experimentation. Clearly, a sound experimental basis is the strongest argument for decision-making. In practice, resources such as manpower, time and capital are limited. If reliable experimental data could be collected automatically and systematically by an appropriately constructed mechanical synthetic device, then more resources could be devoted to the creative aspects of process research and development.

A closed-loop system employing a computer-based simplex optimization algorithm is ideally suited for automated chemical synthesis [1, 2]. This algorithm is an efficient experimental design for optimization experiments [2—7], and can be used for optimizing responses in chemical and analytical systems [2, 7—10].

An automated system for chemical synthesis should optimize, automatically and efficiently, a novel reaction that has been carried out at the bench at least once to produce a sample of the product. Figure 1 shows the components of such a system. The reactor should be completely automated (fill, drain, mix, cool, heat, etc.) with regard to inputs. It should be chemically inert and resemble the geometry of the vessel into which the reaction will ultimately be scaled so that optimized factors can be related to the eventual manufacturing process. The analytical unit should be stable, simple and able to handle a wide variety of reaction products. It should also be capable of standardizing itself and, if possible, computing its own analytical report.

The computer should be programmable in a readily accessible high-level language, and should be easily interfaced with the reactor and analyzer components, as well as capable of providing all the system control logic. It should be flexible and allow for expansion of the design to a network of automated synthesis systems.

The prototype of an automated system of chemical synthesis has been designed that contains many elements of the ideal system. It is built around a PDP-11/34 time-sharing system (RSTS) that is interfaced with the remaining system hardware by a local digital controller (LDC). The system operates on the modified simplex algorithm which is programmed in extended BASIC [11].

The individual components of the system have been tested and are in the process of being interfaced. Model chemical reactions will be studied before chemical syntheses are tried.

EXPERIMENTAL

Reactor unit

The reactor (Fig. 2) is a 100-ml glass (jacketed) model of a commercial

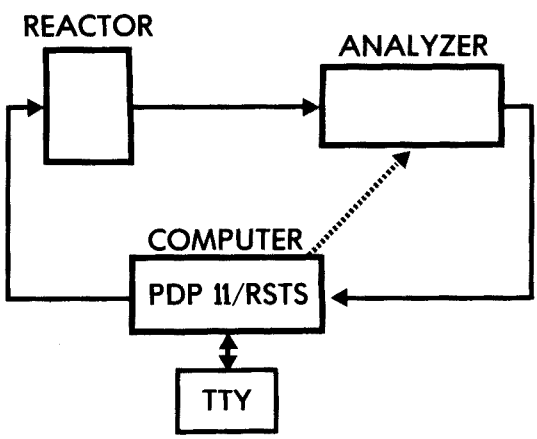


Fig. 1. Closed-loop system for automated chemical synthesis.

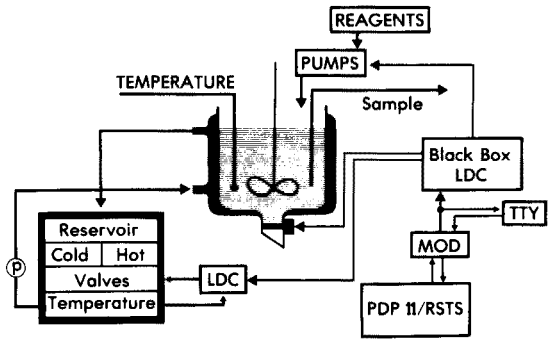


Fig. 2. Reactor unit.

glass-lined reaction vessel used to carry out chemical reactions at atmospheric pressure. The cover of the vessel contains ports for condenser, stirrer, temperature sensor, reagent lines and a sample line. The drain port is sealed by a normally closed Teflon solenoid valve which is under computer control.

The jacket contains silicone heat-exchange fluid (GE 96—50) which is cycled through an automatic constant-temperature control device. The temperature in the jacket can be changed rapidly and then kept constant by balancing streams of hot and cold fluids pumped from their respective constant-temperature baths.

Liquid reagents or solutions of solid reagents are stored in glass or plastic reservoirs and metered into the vessel by four positive displacement pumps (Fluid Metering Inc., Model RHICKE) which have delivery heads adjustable between 0.01 and 0.1 ml per revolution. Each pump is operated by an ordinary a.c. motor connected by a magnetic clutch. Metering is accomplished by counting revolutions, and the total reaction volume can be kept constant by matching the delivery through each pump and maintaining a constant sum of revolutions. Typically, one pump is reserved for the yield-determining reagent and another for solvent to make the sum of revolutions a constant. The reagents come in contact only with glass or Teflon.

The sample line is sealed into the head of the reactor and is made from glass and Teflon capillary tubing until it joins the analytical unit, where other plastic materials can be used. For work at an especially high or low temperature, the sample line is fitted with a heat exchanger so that the sample will always be diluted at ambient temperature.

The stirrer is of constant-speed type, but variable speed control is possible with stepper or other variable-speed motors.

The temperature-sensing device is supported in a glass well. It is a standard semi-conductor junction type with a linear analog output. It can be used to control the reaction temperature or simply to record the reaction temperature.

Analytical unit

The heart of the analytical system (Fig. 3) is a reverse-phase liquid chromatographic column (Waters, C_{18}-Microbondapak) coupled to an Altex Model 110

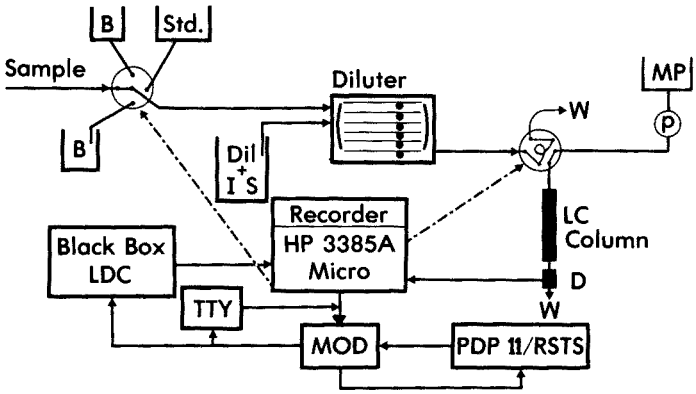


Fig. 3. Analytical unit.

solvent-metering pump and an Altex Model 153 u.v. (254 nm) detector. On the input side of the column is a six-port rotary Teflon valve (Rheodyne Model 50—12 valve with 60°-step pneumatic actuator and Tyna-Myte 4-way pneumatic solenoid) programmed to operate as a four-port valve. The valve directs either a standard for the response, a sample of the reaction mixture, or a solvent blank (or other standby material) to the diluter under control of the local microprocessor (HP 3385A).

The diluter incorporates a Technicon pumping system that dilutes the sample with mobile phase containing an internal standard. The diluted mixture is pumped continuously through the loop of a standard six-port stainless-steel chromatographic injector valve (Rheodyne Model 70—10 valve and 70—01 pneumatic actuator).

On the output side of the detector is a Hewlett-Packard 3385A integrator-microprocessor. The HP 3385A is programmed to operate both the selector valve and the injector valve and performs other programmable functions in the RUN TIME and CHANGE RUN systems. The analysis begins with a computer command to the HP 3385A and ends when the chromatogram is processed, printed and transmitted to the simplex program.

Operation of the system (Fig. 4)

An experiment is started by loading the appropriate reagent and solvent reservoirs and then entering boundary and starting conditions for the factors in the simplex algorithm into the program. Also entered are the retention-time window for the response peak and the fixed concentration of the principal reagent to be used in each experiment (vertex).

The reactor is rinsed by a fill-and-drain cycle of solvent, and is then loaded with reagents for the first experiment (vertex) and adjusted to the reaction temperature. At a calculated time before the end of the experiment, the microprocessor program is started and standard is sent through the diluter and

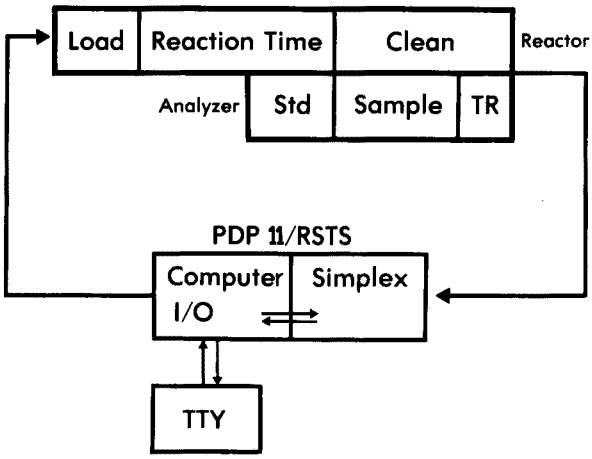


Fig. 4. Reactor and analyzer sequences.

injected onto the column. The response factor for the internal standard is updated and stored by the integrator. The report of this analysis is transmitted to RSTS, but is not used for calculations. The simplex experiment is now over, and RSTS again starts the microprocessor, this time with sample passing through the diluter. The chromatographic analysis this time is accepted by RSTS, and the data are searched for the response peak. When found, the value for the response is stored in the simplex array, and the vessel is brought back to ambient temperature, drained and rinsed. The cycle is ready to be repeated with new inputs from the algorithm. This is repeated sequentially until a pre-determined number of vertices has been explored or until the algorithm senses that an optimum has been found by confirming the response of a vertex retained for $N+1$ simplexes, where N is the number of factors [5].

CONTROL ELECTRONICS

In this system (Fig. 5), commands from a processing program are sent by telephone lines to local hardware. The hardware implements these commands for control of pumps and relays, and the relays in turn control solenoids and a reporting integrator. The program processes the next simplex vertex after receiving a report. The distance between the computer and the laboratory is limited only by the availability of good telephone service.

The processing program is in a time-sharing computer which serves many local terminals. The response time of the program will vary according to the immediate load on the computer and this may introduce more timing errors than can be tolerated. For this reason, hardware in the laboratory stores numerical information from the computer for later use.

In this method the local hardware accepts the bit positions of a string of characters as the control information for the local devices and stores them

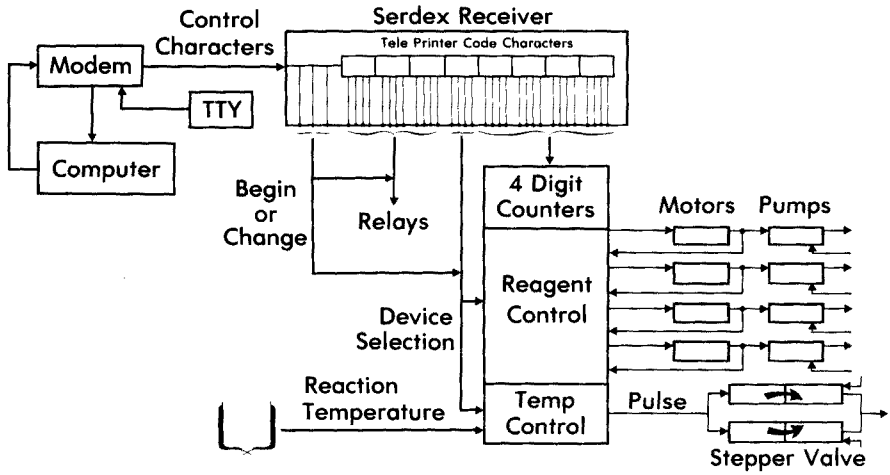


Fig. 5. System control mechanism.

in buffers until the time for starting these devices (Fig. 6). A commercial module is available that combines the input circuits required for interfacing to the telephone modem and those required to store a string of eight characters (Analog Devices, SERDEX Modules, Norwood, Mass.). The limitation is that only four bits of each character are available for the required circuits because the module was designed for the transfer of binary-coded decimal numbers.

A format was chosen in which each of the 12 bits from three characters would control a separate relay, four bits from another character would be decoded to select any one of 16 devices, and the remaining four characters would represent a four-digit decimal number for the selected device (Fig. 6). The state of each of twelve relays can be defined within one command string; however, each string can affect only one number device at a time (Fig. 7). For this reason, the electronics permit loading relays or devices into a local buffer to be held until the moment a new activity begins. Separate GO signals for relays, pumps, and the temperature-controller have been provided.

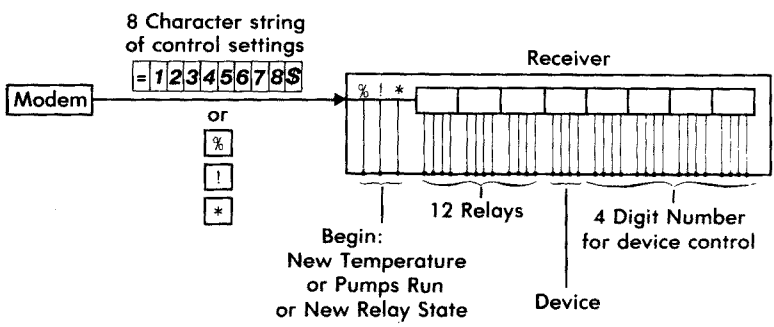


Fig. 6. Input from computer.

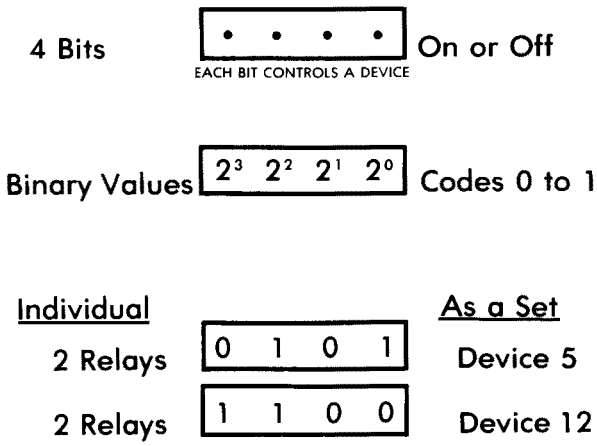


Fig. 7. Control by bit position.

Each pump motor is assigned a device number, a four-digit subtracting counter, and a GO flip-flop (Fig. 8). When a pump is selected, the four decimal digits from the receiver are saved in a buffer which is part of that counter. All the pumps that have been loaded with new information start together when their GO signal is received from the computer. It has been arranged that each pump will send a single pulse to its associated subtracting counter for each 360° revolution of its shaft. When the number in the counter is reduced to zero, the GO flip-flop resets, power is removed from the motor, and a magnetic clutch disengages the motor from the pump.

The subtracting counter design could also be used to count the turns of a screw feeding solid materials or to count the output of a voltage to frequency converter monitoring a strain gauge.

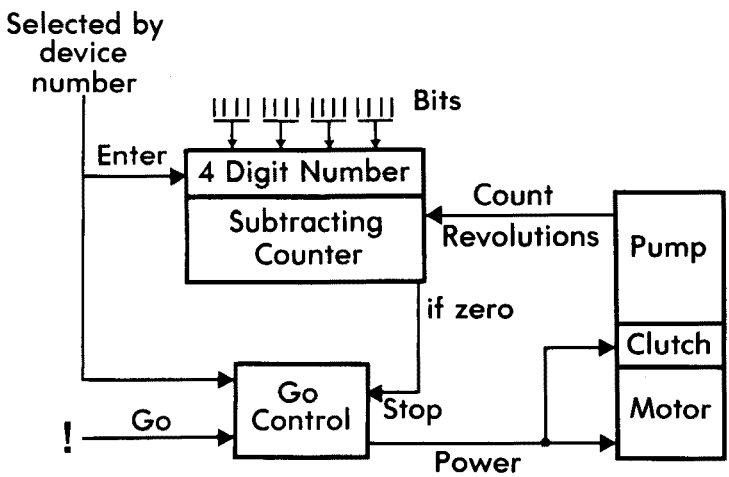


Fig. 8. Pump control.

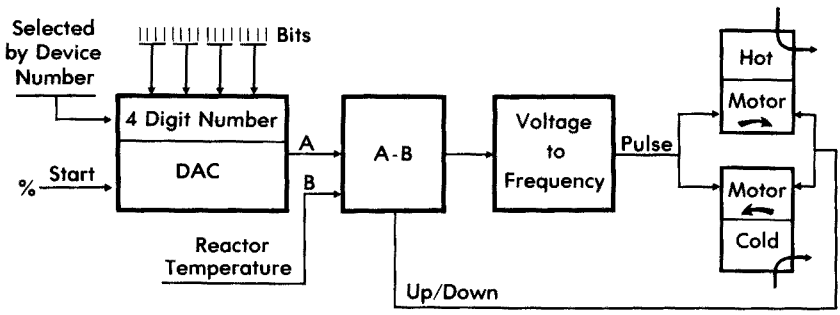


Fig. 9. Temperature control.

One device number is assigned to the temperature-controller (Fig. 9). Three of the four available digits are sent to a buffer which is connected to a D/A converter. The output represents the temperature selected by the simplex program. This reference is compared to a signal representing the temperature in the jacket at the chemical reactor. By varying the mix from hot and cold fluid reservoirs, analog modules control the temperature of the reactor. This is accomplished by subtracting the reference signal from the reactor temperature signal and using the results to derive stepping pulses for motors turning valves which in turn control the mixing process. The motors turn in the direction determined by the sign of the difference signal. The nature of such a feedback system is to change the temperature at the reactor so as to reduce the difference between the reference and temperature signals.

We thank Professor S. N. Deming (University of Houston) for providing a copy of the modified simplex optimization algorithm and for encouragement during this work; we also acknowledge the advice and services of Mr. A. Airey, Mr. G. Gebhard, Dr. K. Kamholz, Mr. F. Wdzieczkowski and Mr. R. Weedon (Smith Kline and French Laboratories).

REFERENCES

1 S. N. Deming and H. L. Pardue, Anal. Chem., 43 (1971) 192.
2 A. S. Olansky, L. R. Parker, Jr., S. L. Morgan and S. N. Deming, Anal. Chim. Acta, 95 (1977) 107.
3 B. H. Carpenter and H. C. Sweeny, Chem. Eng., July 5, 1965, pp. 117—126.
4 J. A. Nelder and R. Mead, Comput. J., 7 (1965) 308.
5 S. N. Deming and S. L. Morgan, Anal. Chem., 45 (1973) 278A.
6 S. L. Morgan and S. N. Deming, Anal. Chem., 46 (1974) 1170.
7 M. W. Routh, P. A. Swartz and M. B. Denton, Anal. Chem., 49 (1977) 1422.
8 D. E. Long, Anal. Chim. Acta, 46 (1969) 193.
9 W. K. Dean, K. J. Heald and S. N. Deming, Science, 189 (1975) 805.
10 G. E. Mieling, R. W. Taylor, L. G. Hargis, J. English and H. L. Pardue, Anal. Chem., 48 (1976) 1686.
11 S. N. Deming, private communication.

Analytica Chimica Acta, 103 (1978) 477—482
Computer Techniques and Optimization

Short Communication

A STATISTICAL APPROACH TO THE BLEND OF INDUCTIVE AND MESOMERIC CONTRIBUTIONS IN DUAL PARAMETER LINEAR FREE ENERGY RELATIONSHIPS

SERGIO CLEMENTI* and FRANCESCO FRINGUELLI
Dipartimento di Chimica, Università di Perugia, Perugia (Italy)
(Received 3rd May 1978)

Dual parameter equations have been used widely over the last two decades for quantitative evaluation of the diverse contributions operating in the effects of substituents linked to aromatic rings. Most equations have been developed for differentiating the relative importance of the inductive and mesomeric contributions. Relevant examples are the approaches suggested by Ehrenson et al. [1] (eqn. 1) and by Swain and Lupton [2]. In the former case, the ratio $\lambda = \rho_I/\rho_R$ was always interpreted as a valid means of defining the blend of the relative contributions of the two effects, but the same is true also for the other equation, where the importance of resonance was given on a percent basis.

$$\log k/k_0 = \rho_I\sigma_I + \rho_R\sigma_R \qquad (1)$$

In previous work [3], some of the statistical limitations implied in these approaches were outlined. In particular, it was concluded that quite often, owing to the small number and inappropriate choice of the data points available for each set of measurements, the differences in the regression parameters caused by variations in the experimental conditions or in the nature of the aromatic substrate were not significant. Moreover, it was shown that the regression parameters of such dual parameter equations were not significantly different from those expected on the basis of the Hammett ρ constant and of the correlations existing between the appropriate sets of σ values. In other words, supporting statistics showed that the numerical values of the regression parameters ρ_I and ρ_R, far from being a good tool for measuring the relative importance of the two types of electronic effect, could simply be calculated for any set of data by using the Hammett ρ constant, because the relative weight of the two effects, given by the relationships between σ, σ_I and σ_R, remains fixed for a fixed set of points. However, in this earlier work, the regression coefficients were tested one at a time. As a chemical meaning is attributed to the ratio λ of the regression coefficients, it would be more appropriate to develop a suitable statistical test of the ratio. A tentative test of the significance of this ratio is discussed in the following paragraphs.

Although it is obvious to a statistician that regression coefficients such as ρ_I and ρ_R cannot be compared directly, it is often difficult for a chemist to realize the high dependence of these coefficients on the number and type of data points (the substituents used), as the independent variables σ_I and σ_R usually cover different ranges. Even worse, the variation of λ in a number of diverse relationships has sometimes been discussed. The correct statistical approach to this problem should involve testing the difference of two ratios, but such a rigorous treatment cannot be used because it is not known whether the error of the ratios is normally distributed.

Selection of significance test

Of the various possible approximate treatments for testing the significance of the ratios between the two regression coefficients simultaneously, it was decided in the present work to study the joint test for the hypothetical values β_1 and β_2 expressed by

$$F_{2,n-3} = [1/2s^2][(b_1-\beta_1)^2 S(x_1^2) + (b_1-\beta_1)(b_2-\beta_2)S(x_1x_2) + (b_2-\beta_2)^2 S(x_2^2)] \quad (2)$$

where the correlation between b_1 and b_2 is taken into consideration [4]. This seems to be the easiest appropriate test to meet the chemical requirements described above. In eqn. (2), the symbols $S(x_i^2)$ and $S(x_ix_j)$ indicate, respectively, the sum of squares and the sum of products of deviations from the mean (deviance and codeviance), whereas s indicates the standard error of estimate.

To simplify further the chemical interpretation of the statistical results, only the drastic limiting case where the two coefficients are not significantly different from each other and therefore their ratio is not significantly different from unity, is considered in this paper. Of course, this would imply that neither of the two contributions could be claimed to prevail on statistically sound bases, and therefore nothing could be said about the variation of the blend of contributions. Accordingly, the restriction $\beta_1 = \beta_2 = (b_1 + b_2)/2$ is applied in eqn. (2).

A further aim of this study was to check from a statistical point of view whether the claimed [1] need for four different sets of substituent resonance constants (σ_R, σ_R^0, σ_R^-, σ_R^+) is really required to improve the significance of the equation in its interpretative potential or is just a device for improving the goodness of fit. Consequently, in all possible sets, the calculations were performed twice, by using σ_R throughout, together with the other appropriate resonance constant.

Results and discussion

The results of the statistical analysis are listed in Table 1. The computations were carried out on a few representatives of the best data sets reported by Ehrenson et al. [1]. The numbers identify the reactions given in ref. [1], and the sets are coded with letters according to the type of reaction examined. In particular, *S* indicates reactions "of the BA type" requiring σ_R as substituent

TABLE 1

Results of the statistical analysis

Regression[a]	With σ_I and σ_R								With σ_I and the appropriate σ_R^0, σ_R^+, σ_R^-					
	n[b]	ρ_I[c]	ρ_R[c]	λ[d]	λ'[c]	F[f]	C.L. for $\lambda = 1$		ρ_I[c]	ρ_R[c]	λ[d]	λ'[e]	F[f]	C.L. for $\lambda = 1$
S1	21*	0.98 ± 0.11	1.01 ± 0.17	1.03	1.40	583	<90							
S2	17*	1.57 ± 0.12	1.28 ± 0.17	0.82	0.56	1573	<99.9							
S8	17*	2.64 ± 0.86	2.27 ± 1.15	0.86	0.64	76	<90							
S26	12	2.49 ± 0.19	2.38 ± 0.22	0.96	0.82	1492	<90							
S28	8	1.78 ± 0.37	1.93 ± 0.58	1.09	0.70	600	<90							
S38	8	9.87 ± 2.04	16.52 ± 2.13	1.67	1.60	708	>99.9							
S40	9	−3.57 ± 0.80	−3.18 ± 1.25	0.89	0.57	270	<90							
Z3	8	0.49 ± 0.14	0.34 ± 0.14	0.70	0.71	136	<95		0.48 ± 0.10	0.44 ± 0.08	0.92	1.20	468	<90
Z12	8	0.62 ± 0.20	0.34 ± 0.30	0.55	0.38	93	<95		0.63 ± 0.19	0.53 ± 0.18	0.85	0.90	226	<90
Z15	21*	−9.84 ± 3.61	−21.1 ± 5.26	2.14	1.47	158	>99.9		−9.94 ± 1.73	−30.8 ± 1.76	3.10	3.03	1464	>99.9
Z19	12*	5.00 ± 1.33	5.17 ± 1.80	1.03	0.76	186	<90		4.90 ± 1.17	7.37 ± 1.14	1.50	1.55	457	<99.9
Z21	18*	5.96 ± 19.8	−93.5 ± 31.6	−15.7	−9.83	16	n.s.[g]		6.76 ± 14.9	−137 ± 16.6	−20.3	−18.2	102	n.s.
Z26	12*	−4.94 ± 2.55	−14.4 ± 4.23	2.92	1.76	121	<99.9		−6.23 ± 2.43	−20.7 ± 2.79	3.33	2.91	304	>99.9
Z27	7	−2.61 ± 4.93	−13.7 ± 6.32	5.26	4.11	41	n.s.		−2.78 ± 7.13	−19.7 ± 6.35	7.09	7.92	46	n.s.
N1	17*	3.41 ± 1.17	3.50 ± 1.47	1.03	0.81	77	<90		3.03 ± 0.14	3.48 ± 0.19	1.15	0.82	5017	<99.9
N11	11*	2.81 ± 1.30	1.89 ± 1.84	0.67	0.47	43	<90		2.37 ± 0.72	1.70 ± 1.30	0.72	0.40	104	<90
N14	11	6.76 ± 2.34	6.39 ± 2.64	0.95	0.84	83	<90		5.96 ± 0.91	5.06 ± 1.36	0.85	0.57	348	<90
N22	13*	6.06 ± 2.08	4.80 ± 2.85	0.79	0.58	70	<90		5.59 ± 0.58	5.04 ± 0.83	0.90	0.63	886	<90
N23	7	6.24 ± 7.98	5.06 ± 11.14	0.81	0.58	12 (97.5)	n.s.		5.65 ± 1.06	6.13 ± 1.49	1.08	0.78	737	<90
N26	16*	1.53 ± 0.51	1.60 ± 0.70	1.05	0.76	98	<90		1.51 ± 0.30	1.56 ± 0.44	1.03	0.70	264	<90
N28	8	−9.99 ± 7.81	−9.06 ± 10.50	0.91	0.67	26 (>99.0)	n.s.		−9.11 ± 3.05	−11.02 ± 3.74	1.21	0.99	214	<90
E1	8	−13.18 ± 4.48	−18.1 ± 6.45	1.38	0.96	200	<90		−11.21 ± 1.48	−9.75 ± 4.10	0.87	0.31	637	<90
E6	14*	5.38 ± 0.81	4.92 ± 1.31	0.91	0.56	438	<90		5.18 ± 0.25	2.76 ± 0.77	0.53	0.17	1810	>99.9
E10	12*	−6.22 ± 1.84	−8.04 ± 2.27	1.29	1.05	134	<90		−5.31 ± 0.40	−4.55 ± 0.89	0.86	0.38	1081	<99.0
E11	13	−4.69 ± 1.36	−6.80 ± 1.31	1.45	1.51	178	<99.0		−4.42 ± 0.40	−4.73 ± 0.56	1.07	0.77	1000	<90
E19	10	−4.92 ± 1.30	−4.75 ± 1.36	0.97	0.92	142	<90		−4.57 ± 0.50	−3.31 ± 0.77	0.72	0.47	491	<99.9
E21	8	−8.08 ± 3.90	−10.0 ± 3.86	1.24	1.26	86	<90		−7.61 ± 1.26	−6.90 ± 1.84	0.91	0.62	408	<90
E22	6	−0.69 ± 0.21	−0.58 ± 0.15	0.84	1.16	392	<90		−0.67 ± 0.10	−0.46 ± 0.09	0.68	0.75	996	<99.9
M1	21*	0.96 ± 0.07	0.30 ± 0.09	0.31	0.21	641	>99.9							
M5	19*	2.22 ± 0.24	0.34 ± 0.34	0.15	0.11	178	n.s.							
M7	17*	2.71 ± 0.28	0.83 ± 0.35	0.30	0.24	294	>99.9							
M11	15*	2.45 ± 0.27	0.70 ± 0.35	0.29	0.22	316	>99.9							
O6	9	2.45 ± 1.40	1.30 ± 1.32	0.53	0.56	22 (>99.0)	n.s.							
O9	14	2.20 ± 0.70	1.37 ± 0.93	0.62	0.47	64	<90							

[a]The intercepts are always not significantly different from zero. The points (0,0,0) are always included in the set. [b]Number of points. An asterisk indicates that the MINIMAL BASIS SET is included. [c]The uncertainties are given as $(t_{0.01,\, n-3} \cdot s_\rho)$ [d]$\lambda \equiv \rho_R/\rho_I$. [e]Ratio of the standardized ρ_R/ρ_I, to allow for the different ranges covered by σ_I, σ_R: $b'_i = b_i (S(x_i^2)/S(y^2))^{\frac{1}{2}}$. [f]$F$-test: the confidence levels are always

constants, Z reactions requiring σ_R^0, N reactions requiring σ_R^-, and E reactions requiring σ_R^+, all referring to *para* reactivities, whilst M and O indicate data sets referring to reactivities at *meta* and *ortho* positions requiring σ_m and σ_o, respectively.

The analyses were carried out on a desk computer to calculated the F-tests for the whole regression and for the null hypothesis $\lambda = 1$ (eqn. 2). The confidence levels for the former test (not shown in Table 1) were always higher than 99%, except for N23. In the latter case the chemical hypothesis is met by confidence levels lower than 99% (null hypothesis accepted) and is considered not to be met by confidence levels higher than 99% (null hypothesis rejected). In a few cases, it was meaningless to test the null hypothesis because the standard error of one or both the regression coefficients was greater than the coefficient itself.

The main features arising from the results can be summarized as follows.
(a) The use of a single set of substituent constants (σ_I and σ_R) is shown to be a statistically adequate model, as the F-tests for the whole regression are good. As this treatment also allows all the data to be compared, independently of the type of reaction considered, there seems to be no need for four different sets of resonance constants.
(b) Within the 34 cases examined by σ_I and σ_R, λ is significantly different from unity in five sets only. In two cases (S38 and Z15) , both of which refer to spectral data, the resonance effect prevails, whereas in the other three (M1, M7, M11), which refer to *meta* reactivities, the inductive effect prevails.
(c) In five further cases, λ is not significant as the experimental data depend on one variable only (Z21, Z27, N28, M5, O6). In these cases the use of a single-parameter, Hammett-type relationship seems more appropriate.
(d) In 21 cases, the appropriate separate sets for the resonance constant (σ_R^0, σ_R^+, σ_R^-) were used. Compared with the normal treatment with σ_R, this approach obviously improves the goodness of fit and therefore the regression in general, but only in four series does this affect the significance of λ. This appears to be too slight an improvement to justify preventing the comparison of different types of reactions.
(e) When the standardized coefficients are compared, to allow for the different ranges covered by the variables, the ratio is always smaller, as a consequence of the peculiar distribution of the pairs σ_I and σ_R (see Fig. 1). The actual magnitude of the difference depends again on the particular substituents used in each set.

In conclusion, this report shows that it is statistically possible to find values of λ significantly different from unity, but this happens in very few cases, although selected among the best data series available in ref. [1] : eight sets only out of the 55 examined in total.

The conditions required to obtain a positive significance test are that the values for ρ_I and ρ_R should not be too similar, and that the regression should be excellent. Excellent regression in turn requires the inclusion of a "minimal basis set" and a very high goodness of fit. A minimal basis set is the smallest

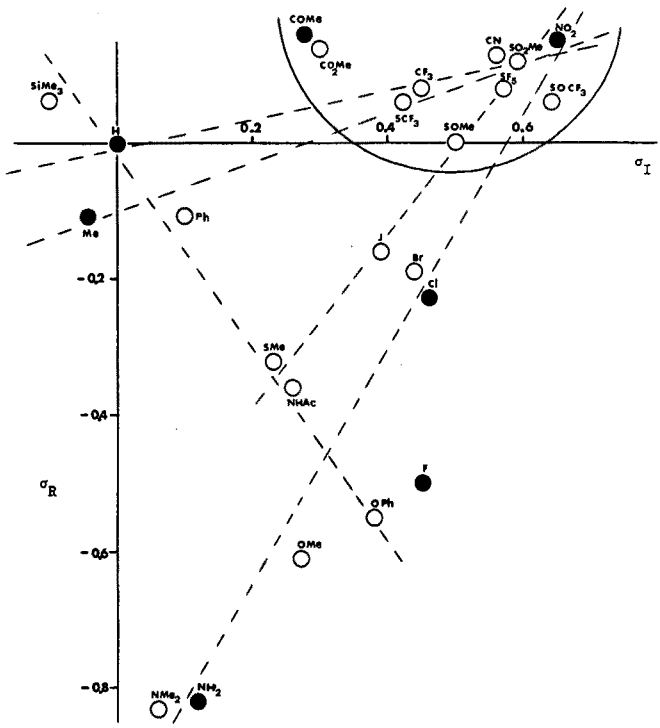


Fig. 1. Plot of σ_I vs. σ_R for the commonest 24 substituents. Note that about a half the points are included in a small area. The dashed lines indicate multicollinearity. The filled circles indicate a possible minimal basis set (see text).

appropriate set of data points: a limit number (say at least seven) and a suitable variety of points (to cover a large area on the σ_I vs. σ_R plane) in the absence of multicollinearity (see Fig. 1). The need for the inclusion of a minimal basis set has already been pointed out [1]; this report shows how, if just one of the conditions required is not met, the significance level decreases steeply.

The conditions required for obtaining statistically significant results so that variations of the blend of inductive and mesomeric contributions can be evaluated are therefore quite severe. Hence, it appears that a large quantity of the literature data on this subject is inadequate for any realistic discussion on the relative importance of the two effects. However, it appears inappropriate to face the problem of substituent effects with such simple statistical tools. Some recent work [5, 6] suggests that more sophisticated statistical approaches can be used successfully to describe chemical reactivity in terms of substituent effects.

The authors thank Prof. C. Scala (Siena) and Dr. G. Savelli (Perugia) for helpful discussions, and C. N. R. (Rome) for financial support.

REFERENCES

1 S. Ehrenson, R. T. C. Brownlee and R. W. Taft, Progr. Phys. Org. Chem., 10 (1973) 1.
2 C. G. Swain and E. C. Lupton, J. Am. Chem. Soc., 90 (1968) 4328.
3 S. Clementi, F. Fringuelli, P. Linda and G. Savelli, Gazz. Chim. Ital., 105 (1975) 281.
4 A. Hald, Statistical Theory with Engineering Applications, Wiley-Toppan, Tokyo, 1952, p. 631.
5 S. Wold and M. Sjöström, Chem. Scripta, 2 (1972) 42; 6 (1974) 114; 9 (1976) 200.
6 S. Wold, Pattern Recognition, 8 (1976) 127.

Analytica Chimica Acta, 103 (1978) 483—487
Computer Techniques and Optimization

Short Communication

HIERARCHICAL LABORATORY AUTOMATION AND FILING SYSTEM FOR PHARMACEUTICAL QUALITY CONTROL

R. van WIJK

Quality Control Department, Organon International, Postbus 20, Oss (The Netherlands)

(Received 3rd May 1978)

Pharmaceutical manufacture is, among other things, governed by severe regulations from national and international authorities; this is reflected in the many prescriptions and other documents that are required. Quality control, involved in many stages of the manufacturing process, is responsible for a large part of this paperwork. The introduction of the directives for Good Manufacturing Practice and more recently Good Laboratory Practice has initiated a general rethinking on data-collection and filing in quality control [1—5]. In 1976 this company investigated whether a computer could be helpful in quality control activities. This resulted in a rough draft of four systems, of which the first is almost realized and is about to be implemented in full. In the present report, some aspects of the general philosophy of computerization will be presented as well as more specific information on each of the four systems.

Quality control activities

The Quality Control Department is responsible for:

taking samples from all incoming raw materials, intermediates, and finished products in a prescribed statistical way, details depending on the lot or batch size and the method of packing;

chemical, microbiological and pharmacological control analyses on the samples in accordance with prescribed methods;

release or rejection of the lots or batches in terms of the results of the analyses etc. and the product specifications;

handling user complaints, where all relevant information on the production batch, raw materials used, etc. is needed quickly.

In general all the above activities are standardized procedures; all the information must be fully recorded (including the identity of the persons performing the activities) and a quick retrieval (cross references) of the recorded information is needed. For a variety of reasons there is an almost continuous stream of updating of the various prescriptions.

Targets for the introduction of (computer) systems

Without increasing manpower, the main targets are to reduce the time needed to perform all the release procedures, either in general or for a number of specific products; to improve the reliability of the data and eliminate the need to copy the data repetitiously through capturing the data at the source and evaluating the data statistically to enhance release judgement, production performance, analysis performance, and knowledge of products and processing; to obtain easily retrievable documentation (filed over a period of 10 y); and to obtain an easy updating and distribution of basic documents (prescriptions).

Development and implementation philosophy

The (sub) systems should be designed as logical entities with the boundaries chosen at points of minimal interactions (transactions) between the systems and so that the systems can be realized independently of each other. The systems are arranged in order such that the first system arranges the work load, optimizes the work sequence and in general makes the monitoring of the work flow easier. In the second system the data are filed, evaluated, and interpreted. Until this is realized, little advantage can be gained from the third system which will speed up the actual analyses through automation and computerization in the laboratory. An organization that is inexperienced in computerized systems should adapt itself gradually to this through occasional laboratory automation or by testing larger systems on a small part of the project. Care must be taken not to use untried systems (hardware or software) through which the reliability of the quality control work could be adversely affected.

Technical resources

For laboratory automation a hierarchical approach (Fig. 1), for which the general tools are already available in the company, is used. Data capture at the work place will be realized through microcomputers coupled to an instrument or experiment. Preferably the instrument or experiment should be controlled through feedback from the microcomputer. Future instruments will be controlled by built-in microprocessors which will also make the calculations (data reduction). However, extra information must be keyed in.

For further filing, evaluating, interpreting, etc., the results from the analyses are either directly, or via buffer storage, transferred to a local time-sharing mini-computer system. Some calculations can be performed there and some reporting can take place. If necessary for filing or complex calculations, it is possible to connect the local time-sharing computer to the number-cruncher from the parent company, either on a several times a day service by remote job entry or almost right away via a time-sharing option. An Akzo microprocessor for laboratory automation is available, based on Intel sbc 80/10 chips, equipped with a Honeywell keyboard, a Thorn Automation strip printer, and a Burroughs display.

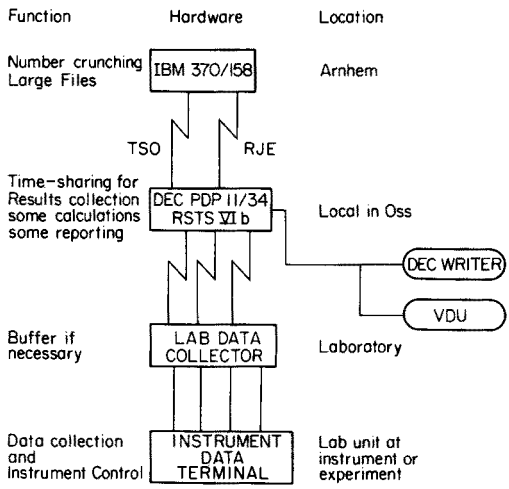


Fig. 1. Hierarchical system. TSO, time-sharing option; RJE, remote job entry; RSTS, resource sharing time-sharing operating system version VI b for PDP 11 (DEC; Digital Equipment Corporation); WRITER TERMINAL type LA 36 (DEC); VDU, visual display unit type ADM 1 A (Lear Siegler Inc.).

Planning and control system

This system mainly registers the workload presented to the Quality Control Department, for supporting the planning in the analytical units and controlling the throughput of the release procedures.

On the local time-sharing system via preformatted dialogues, planned dates for production of a batch or delivery of a lot, subsequent dates for availability of samples, and required release dates are supplied by the production planners. The laboratory units give planning dates for the start and finish of their activities. As soon as a planned date becomes a reality, this date (now asterisked) replaces the planned date. Also via dialogues, several people can obtain lists sorted to some element or a combination of elements that best suits their needs for planning or control purposes.

In the system the information can be read on all terminals, but the entering or changing of dates is restricted to those responsible for a specific date. As a result of a provisional test on a selected 10% of the workload, final specifications have already been drawn up. The system should be fully implemented in 1978.

Release system

This system is a data base for the results from analyses and is to support the release decision and to make the information available for use at several places in the Production, Research, and Quality Control Departments. At present, the analyst reports the analytical results on a form. In this second system the results will be keyed in. In the future, through the third system,

data collection will be automated as far as possible. With the help of statistical programs adapted to the specific determination, the analytical results will be weighed against the required specifications and against the product history. As a result of this evaluation the laboratory unit may decide to repeat the determination to ensure that any discrepancy found did not occur accidentally and to confirm the results reported. The system will check if all laboratory units have reported and confirmed their results for a specific release and print the required documents that will be signed for release or rejection depending on the decision of the quality control manager. When this decision is keyed in, several documents will be generated, and the product data base will be up-dated with the new results.

Laboratory data system

This is mainly a data collection system and is preferably connected directly to instruments. Some additional data (batch number, identity of analyst, etc.) have to be entered in a simple but reliable way. The use of a reading pen or identity cards is under consideration. In principle, the Instrument Data Terminal will reduce the measured values to results which will be transferred to the local time-sharing system directly or via buffer storage (laboratory data collector) depending on data rates, time-critical determinations, etc.

All relevant data will be printed in order to check keyed-in identifications, measured values, etc., and reconstruct analytical results independently of the system (security). The print strip, attached to the batch-instruction forms, will replace both the present forms and the analyst's note-book.

At present, there are two individual automated tests, which serve also as pilot studies for subsequent automation, because several practical and organizational problems had to be solved. The first system is a dual Varian 240 Chromatography Data System to which about 65 chromatographic detectors (gas and liquid etc.) are connected, including those from the Research Department. With the recently delivered Class V software it will be possible to connect the systems to the local time-sharing computer for further use of the data. In the other system (a Wang 600 desk computer with a dual tape unit, connected to a Mettler electronic balance and a spectrophotometer equipped with a sampler) the weight distributions of series of tablets are determined as well as the contents of the individual tablets. The statistically calculated results are printed with the relevant additional text in the form of a report per batch.

Central documentation systems

The Quality Control Department is responsible for providing forms, prescriptions etc. Technical details are supplied by the laboratory managers, etc., and the central group handles the lay-out, issue, up-dating, filing, reproduction distribution, etc. To assist in these activities there are now available 2 Redactron (Burroughs) Redactor I word processors with magnetic cards. In due course, all the existing pages will be on magnetic memory. In this way, initiating new pages and up-dating old ones will be easier.

In principle, the word processor can be connected to a computer via a transmission device. Within the systems the best way of making available the latest issue of a document where and when it is needed is under consideration. The word processor can also be used for the print-out of well-presented completed documents, e.g. certificates.

It is hoped to have the set of systems implemented in 1981. Continuous improvements should then be possible, e.g. through improved analytical instrumentation. It is clear that the systems are basically administrative tools. They are being introduced because of the severe regulations imposed by the authorities. The aim is to improve documentation and to reduce the burden of paperwork in the laboratories.

REFERENCES

1 A. R. Clare and J. C. Hall, Proc. Soc. Anal. Chem., (1973) 295.
2 H. Feltkamp, Pharm. Ind., 34 (1972) 420.
3 D. G. van Doren, Proc. Digital Eq. Comp. Users Soc., (1975) 235.
4 L. Ehrhardt, Pharm. Ind., 36 (1974) 493.
5 H. Ganshirt, Pharm. Ind., 36 (1974) 478.

Analytica Chimica Acta, 103 (1978) 489—494
Computer Techniques and Optimization

Short Communication

A MICROCOMPUTER-CONTROLLED PHOTOMETRIC ANALYZER
Application to the Determination of Lidocaine

LEIF ANDERSSON, ANDERS GRANÉLI* and MATS STRANDBERG**
Department of Analytical Chemistry, University of Göteborg, Fack, S-402 20 Göteborg (Sweden)

(Received 1st June 1978)

When an analytical instrument is computerized, the best solution is not always achieved by simply repeating the manual procedure with the computer. In an automatic titrator, for instance, the computer may be used to control one or several motor syringe burets to deliver reagents into a titration vessel, to measure the signal from some sensor such as an electrode or a photometer, and finally to calculate the equivalence volume. However, a more flexible instrument may be achieved by simply using the buret as titration vessel, drawing solutions into it. Such a system has been used in a minicomputer-controlled phototitrator [1]. In many instances, the computing power and speed of minicomputers is not required to automate titrators; a microcomputer will do instead. The use of a microcomputer to control a titrator and its application to the photometric titration of lidocaine in acetic acid medium are described below.

Experimental

The microcomputer system. The microcomputer used was intended for a more or less general laboratory system, and therefore required an A/D converter with sufficient range to measure signals from a variety of detectors, as well as a simple means of controlling instruments and sufficient memory for easy reprogramming for new applications. The system chosen was a commercial single board computer (DIM 1001; A/S Mycron) with an INTEL 8080A processor. In addition to the processor and the logics associated with it, the board contains 1 kbyte static RAM, sockets for 2 kbyte PROM, priority encoder for seven interrupt levels and one UART for serial input/output [2].

To complete the system, two more cards were added, one memory card and one data acquisition card. The memory card contains 2 kbyte static RAM and 24 programmable input/output lines, eight of which are used to control reed relays, four of them for driving two stepper motors and twelve for

**Present address: Astra Pharmaceuticals AB, S-151 85 Södertälje (Sweden)

digital input/output. The data acquisition card consists of a 10-bit A/D converter (Analog Devices AD 7570) with a ± 10 V range, sample and hold circuits, programmable amplification up to 256 times and an 8-channel, single-ended analog multiplexer. All input/output is memory-mapped which considerably simplifies programming. A schematic diagram of the computer is shown in Fig. 1 and a programming example for reading the A/D converter is shown in Table 1. The amplification is set by writing the factor into address 1880H (H = hexadecimal). Conversion is then started by writing anything into address (1800 + X)H where X represents the channel number. After a 10-μs settling time and a 20-μs conversion time, the eight least significant bits may be read from address 1801H and the two most significant bits from 1802H. The maximum conversion rate of the A/D is approximately 20 kHz.

Titrator. The titrator is a modified version of the one described previously [1]. A schematic diagram is shown in Fig. 2. The titrator consists of a 20-ml motor syringe buret (Metrohm Dosimat E412) where the ordinary two-way stopcock has been replaced by a Teflon lid with a single outlet/inlet connected to a 6-way motorized rotating valve (Altex Scientific Inc.). The valve is connected to five different bottles containing either sample or reagents. Solutions can thus be sucked into the titrator from any of the bottles through the valve. The sixth connection of the valve is used to empty the titrator. The a.c. motor of the buret was replaced by a four-phase stepper motor (Philips 9904 112 06001). Mixing of solutions in the titrator is achieved by rotating a Teflon-clad magnetic ball which fits into a hemisphere in the titrator piston and a hemisphere in the titrator lid (cf Fig. 2). Thus the titrator is emptied with the piston in the top position, the dead volume being negligible. In the present application the instrument is used as a phototitrator. Transmittance is measured across the buret by using a fiber-optics light guide photometer (Brinkman PC/600). Thus the buret acts simultaneously as a cuvette, a volumetric measuring device and a titration vessel.

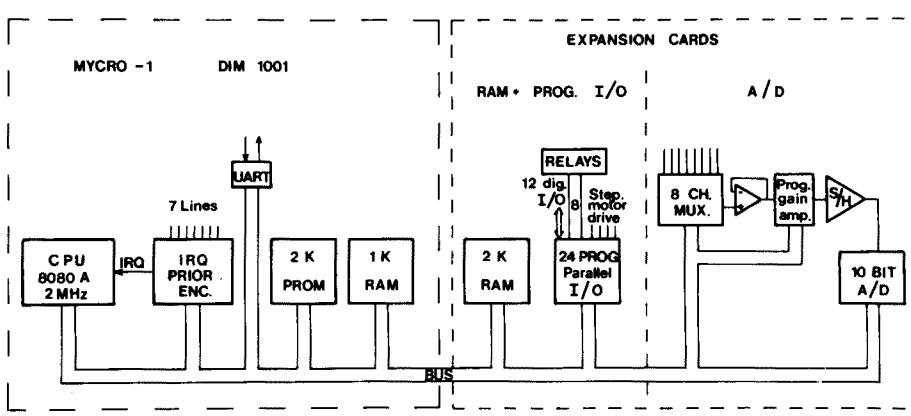


Fig. 1. Schematic diagram of the microcomputer showing the processor card and the two expansion cards.

TABLE 1

Subroutine for reading the A/D converter

```
LOC   OBJ      SEQ      SOURCE STATEMENT
                  1     ;THIS SUBROUTINE READS THE ADC 16 TIMES
                  2     ;AND RETURNS THE MEAN VALUE IN D,E REG.
                  3     ;IT CONNECTS THE ADC TO THE CHANNEL
                  4     ;SPECIFIED IN C REG. AND SETS AMP. TO
                  5     ;GAIN SPECIFIED IN B REG.
                  6     ;ALL REG. DESTROYED
                  7     ;
                  8     ;
2500              9     ORG   2500H
                 10     ASEG
2500  110000   11 AD:   LXI   D,0
2503  218018     12     LXI   H, 1880H  ;ADDRESS TO PROG. GAIN AMP.
2506  70         13     MOV   M,B       ;SET GAIN
2507  0610       14     MVI   B,10H     ;LOOP COUNTER
2509  2618    15 AD1:   MVI   H,18H     ;ADDRESS TO MUX AND ADC
250B  69         16     MOV   L,C
250C  77         17     MOV   M,A       ;DUMMY WRITE SETS CHANNEL
                 18                     ;AND STARTS ADC
250D  3E09       19     MVI   A,9H      ;DELAY FOR ADC BUSY
250F  3D      20 WAIT:  DCR   A
2510  C20F25     21     JNZ   WAIT
2513  2A0118     22     LHLD  1801H     ;READ ADC
2516  7C         23     MOV   A,H       ;MASK OFF 6 MSB
2517  E603       24     ANI   3H
2519  67         25     MOV   H,A
251A  19         26     DAD   D         ;ADD TO PREVIOUS READINGS
251B  EB         27     XCHG            ;STORE IN D,E
251C  05         28     DCR   B
251D  C20925     29     JNZ   AD1       ;16 READINGS?
2520  0604       30     MVI   B,4H      ;SHIFT COUNTER
2522  AF      31 DIV:   XRA   A
2523  7A         32     MOV   A,D       ;16 BIT SHIFT RIGHT. ZEROES
2524  1F         33     RAR             ;ARE SHIFTED INTO THE
2525  57         34     MOV   D,A       ;HIGH ORDER BITS
2526  7B         35     MOV   A,E
2527  1F         36     RAR
2528  5F         37     MOV   E,A
2529  05         38     DCR   B
252A  C22225     39     JNZ   DIV       ;FINISHED?
252D  C9         40     RET
                 41     END
```

Application to the determination of lidocaine. The functioning of the instrument is illustrated by the photometric titration of lidocaine (a local anaesthetic manufactured by Astra Pharmaceuticals under the name xylocaine). It is a weak base with a pK_a value of 7.86 in water [3]. However, it may be titrated with perchloric acid in acetic acid—dioxane medium in

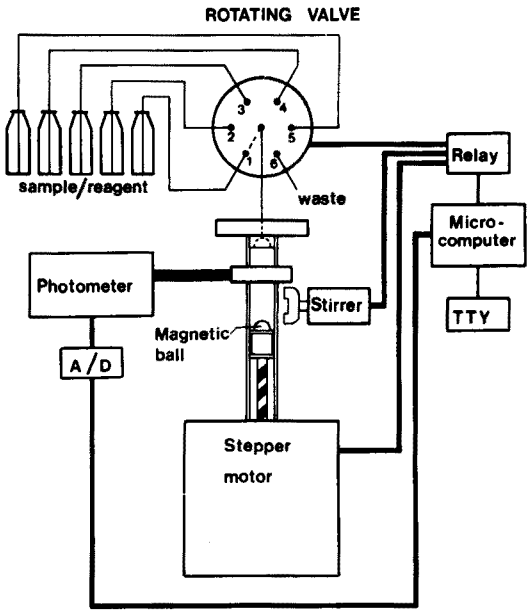
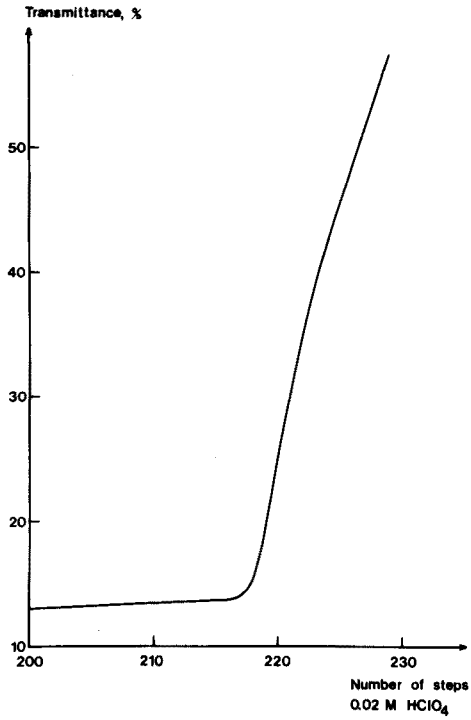


Fig. 2. Schematic diagram of the photometric analyzer.

the presence of malachite green indicator. The method is a modification of that used by Astra Pharmaceuticals to control the purity of the end-product in the lidocaine synthesis. The main difference is that malachite green is used instead of crystal violet as indicator [4].

Approximately 0.5 g of the lidocaine sample was dissolved in 100 ml of anhydrous acetic acid. A 0.02 M perchloric acid (standardized against potassium hydrogenphthalate) and a 0.1% indicator solution in anhydrous acetic acid were also prepared. These solutions together with dioxane were connected to four different inlets of the valve. Since the actual dimensions of the volume measurements are not essential for calculation of the sample concentration, it is practical to use the number of steps of the stepper motor as the unit. In this titration 200 steps of sample, 500 steps of dioxane and 2 steps of indicator were found to be suitable (1 step is approximately 0.02 ml). In order to reach the vicinity of the equivalence point rapidly, 200 steps of titrant were added before commencement of the photometric measurements. The transmittance was then measured for every additional step of titrant. As the titration curve has a pronounced break point (Fig. 3), the equivalence volume is easily determined by comparing consecutive transmittance values. The procedure is outlined in the flow chart (Fig. 4). Any water present in the solvent is titrated simultaneously with the lidocaine. Therefore the water content must be determined and corrected for by a separate titration of pure solvent.

The program consists of subroutines for controlling valve, stepper motor



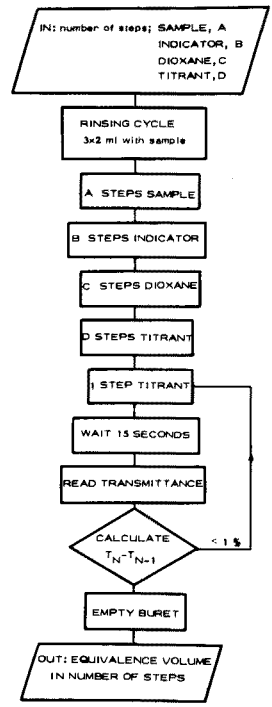


Fig. 3. Titration curve of lidocaine titrated with perchloric acid (malachite green indicator).

Fig. 4. Flow chart of the titration program.

and stirrer motor respectively, for reading the A/D converter (cf. Table 1), generating delays, etc. These modules can easily be linked together to produce a suitable program for most titrations.

Results and conclusions

To evaluate the precision of the volumetric measurements of the titrator, four 100-step portions of distilled water at 22.5°C were dispensed into beakers and weighed. This gave a relative standard deviation of 0.03% which agreed well with earlier values [1, 5]. Filling the titrator with 20 ml of solution required about 1000 steps by the motor, giving a resolution in the volumetric measurement of ca. 0.02 ml, compared to 0.1 ml with the original motor. When a stepper motor is used, the speed and direction of the piston can easily be altered by the computer.

The percentage (w/w) of lidocaine of the original sample was calculated from the mean equivalence volume (corrected for water content) of ten consecutive titrations. The lidocaine content was found to be 99.8% (relative

standard deviation, 0.6%). Considering the simplicity of the evaluation method, this result is satisfactory.

Since the titrator is completely closed, it is particularly useful in titrations with hazardous chemicals, such as drugs and solvents, or in methods where air might interfere, as in the Karl Fischer titration. The lidocaine titration requires only integer arithmetic in order to determine the equivalence volume. However, if the computer has floating point arithmetic, more elaborate evaluations, such as linear regression or least-squares curve fitting, may be used. The instrument might then be used in any photometric analysis, not only titrations, the main drawback being that programming is done in assembler. An even more flexible and powerful instrument could be obtained if the computer were equipped with a PROM-resident BASIC interpreter [6]. Such a system can easily be programmed by any chemist to perform photometric analysis by titration, calibration or standard addition, to prepare suitable standard solutions from stock solutions or even to act as an extraction unit in analytical procedures involving two-phase equilibria.

The authors are indebted to Professor David Dyrssen for valuable discussions. Grants from the Swedish Natural Sciences Research Council are gratefully acknowledged.

REFERENCES

1 A. Granéli and T. Anfält, Anal. Chim. Acta, 91 (1977) 175.
2 MYCRO-1, Dim 1001 Computer module users guide.
3 K. Gustavii, P-A. Johansson and A. Brändström, Acta Pharm. Suec., 13 (1976) 391.
4 E. Bishop, Indicators, Pergamon Press, Oxford, 1972, p. 185.
5 D. Jagner and K. Arén, Anal. Chim. Acta, 52 (1970) 491.
6 R. Langer and T. Dugan, Electronics, 51(8) (1978) 119.

ANALYTICA CHIMICA ACTA, VOL. 103 (1978)
(Computer Techniques and Optimization, Vol. 2, No. 4)

AUTHOR INDEX

Mathematics and Computers in Simulation

Subscription Information

1979 Volume 21
One volume in six issues
Subscription prices for 1979:
Institutional: US $64.50/Dfl. 145.00 (postage included)
Personal: US $31.50/Dfl. 71.00 (postage included)
Personal subscriptions are available from the publisher only.

Aims and Scope

The aim of the journal is to provide an international forum for the dissemination of up-to-date information in the field of Computer Simulation of Systems.

MATHEMATICS AND COMPUTERS IN SIMULATION is the official organ of the International Association of Mathematics and Computers in Simulation (formerly AICA).

Topics covered by the journal include mathematical tools in the foundations of Systems Modelling, in specific applications to science and engineering, and in the supporting analysis of numerical methods. They also include considerations about computer hardware for simulation (digital, analog and hybrid) and about special software and compilers. Finally, the journal publishes articles concerned with the general philosophy of systems simulation and its impact on disciplinary and interdisciplinary research. The journal also includes a Bibliography and Book Review Section.

Recent Papers:

A Continuous-System Simulation Language Designed for LSI Economics *(D. M. Auslander)*
Spatially Discrete Models of Counter-Current Mass Transport for Application to the Kidney *(G. M. Saidel, M. A. Knepper and P. S. Chandhoke)*
Hybrid Simulation of Adaptive Open Loop Control for Parabolic Systems *(M. Amouroux, A. El Jai, J. P. Babary and J. P. Gouyon)*
New Methods for Evaluating the Validity of the Results of Mathematical Computations *(J. Vignes)*
A Kalman Filter Type of Extension to a Deterministic Gradient Technique for Parameter Estimation *(A. P. Roberts and M. Smith)*
Calculation of Demagnetizing Field by Means of FFT *(Yih-O-Tu)*
Discrete Event Simulation Modelling of Computer Systems for Performance Evaluation *(J. Leroudier and M. Parent)*
New Method for Analyzing Static Converter Behaviour Using Global Analogue Simulation *(M. Metz, H. Foch and B. Trannoy)*
Simulation Analogique D'un Convertisseur Electrique Statique a l'Aide D'un Calculateur Reversible *(T. Maurin and C. Sol)*

FREE specimen copies are available upon request from the publisher.

north-holland

P.O. BOX 211 / 1000 AE AMSTERDAM / THE NETHERLANDS

IN THE U.S.A. AND CANADA: 52 VANDERBILT AVENUE / NEW YORK, N.Y. 10017

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

6003 NHB

Printed in The Netherlands

VOL. 103, No. 4 ANALYTICA CHIMICA ACTA December 15, 1978

(Computer Techniques and Optimization, Vol. 2, No. 4)

CONTENTS

International Conference on Computers and Optimization in Analytical Chemistry, Amsterdam, April 5–7, 1978