ol. 112 No. 3 September 15, 1979 ISSN 0378-4304
omputer Techniques and Optimization, Vol. 3 No. 3)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

EDITOR
J. T. CLERC (Bern, Switzerland)
Associate Editor
E. ZIEGLER (Mülheim, Germany)
Editorial Advisers
R. E. Dessy, Blacksburg, Va.
J. W. Frazer, Livermore, Calif.
H. Günzler, Ludwigshafen
S. R. Heller, Washington, D.C.
J. F. K. Huber, Vienna
T. L. Isenhour, Chapel Hill, N.C.
P. C. Jurs, University Park, Pa.
M. Kneidl, Munich
D. L. Massart, Sint Genesius-Rhode
H. C. Smit, Amsterdam

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 1979 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	104/1	104/2	105	106/1	106/2	107	108	109/1	109/2	110/1	110/2	111
Section on Computer Techniques and Optimization			112/1			112/2			112/3			112/4

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 9 volumes in 1979. The subscription for 1979 (Vols. 104–112) is Dfl. 1179.00 plus Dfl. 135.00 (postage) (total approx. U.S. $641.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 112) is Dfl. 131.00 plus Dfl. 15.00 (postage) (total approx. U.S. $71.00). 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. 1–103) except Vols. 23 and 28 are available at Dfl. 150.00 (U.S. $73.20), plus Dfl. 10.00 (U.S. $4.90) 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.

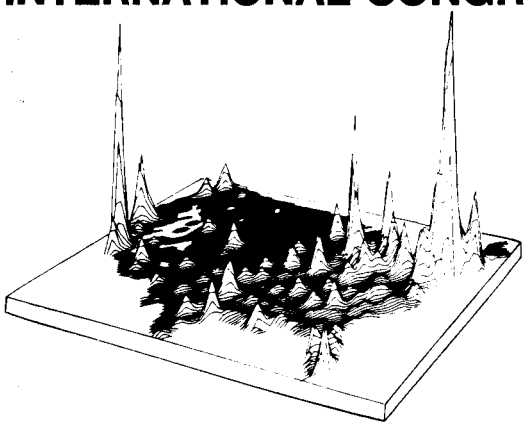
GEORGE WASHINGTON UNIVERSITY
School of Engineering and Applied Science

Announces its

CONTINUING ADVANCED ENGINEERING PROGRAM IN BERLIN

at the new

INTERNATIONAL CONGRESS CENTER



1980 Series of 75 Seminars

Featuring:
Computers
Electronics
Communications
Data Management
Reliability
Structures
Mathematical Methods
Engineering and
Cost Management
Energy

A program for engineers, scientists, and managers in industry and government which stresses practical applications. These seminars of 3 to 5 days in length can benefit you in your job, and aid your career advancement.

Please send me the brochures of the
CONTINUING ADVANCED ENGINEERING PROGRAM

Name_____
First Middle Last

Title_____

Organization_____

Address_____

City and Zone_____

Country_____

Telephone_____

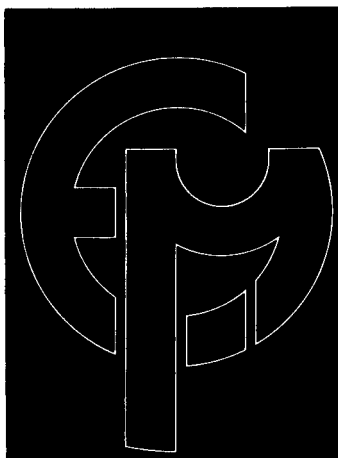
200

PLEASE CHECK YOUR AREAS OF INTEREST

Computers & Data Management _____
Electronics & Communications _____
Reliability & Safety _____
Medical Engineering _____
Structures & Materials _____
Math Methods _____
Engineering & Cost Management _____
Environment _____
Energy _____
Other _____

For further details send this panel to the address shown below.

ADDRESS TO:
Continuing Engineering Education
George Washington University
Washington, D.C. 20052 U.S.A.



EUROMICRO JOURNAL

The international journal of microprocessing and microprogramming

Scope:

The main purpose of the EUROMICRO JOURNAL is to facilitate the international flow of information in the field of micro-processing and micro-programming. In addition to reports on recent research and technological progress, the journal contains comprehensive surveys, state-of-the-art reports, short communications and announcements pertinent to the micro field.

Theory, design, research, languages, simulation, emulation, teaching aids and evaluation/diagnostic methods are all within the scope of this journal.

Topics relating to the following subjects are covered: micro-programming, micro-processing, micro-processor systems and networks, distributed computing, MSI/LSI components, computer structures, modular systems, integrated hardware/software design and micro-architecture of computer systems.

The EUROMICRO JOURNAL is the official organ of the European Association for Microprocessing and Microprogramming, 313 Rue Lecourbe, 75015 Paris, France. Since January, 1976 the Journal has been published and distributed by North-Holland both for the EUROMICRO membership and regular subscribers.

Editorial Organization:

EUROMICRO's policies are set by its Board of Directors, which also acts as the Board of Editors for the Journal. It includes National Correspondents for the countries active in the Association, with a balance between university and industry representatives.

Editors:

Editor: Giovanna Sami, Milan
Technical Editor: Helmut Berndt, Munich
Section Editors - Book Reviews: Lorenzo Mezzalira, Milan and Ole Caprani, Copenhagen.
Section Editor - Calendar of Events: Reiner W. Hartenstein, Kaiserslautern
Section Editor - Standards: Giovanna Sami, Milan

Recent Papers:

Volume 4, no. 4

Improvements of the multiprocessing capabilities of micro-processor busses *(W. Mahr and R. Patzelt, Austria).* Debugging aids for real time microprocessor systems *(P. Lenders and J. Tiberghien, Belgium).* GPCA: a general purpose cross-assembler *(Sergio Antoy, Franco Cordano, Franco Serio and Tullio Vernazza, Italy).* Parallel algorithms for solving matrix-oriented problems on the three-dimensional computer *(William A. Hanna, USA).* Microinformatics and work organization - a case study *(C. Ciborra, G. Gasbarri and P. Maggiolini, Italy).* Some microprogramming and microprocessor activities in Sweden *(Gunnar Carlstedt, Harold W. Lawson, Jr. and Lars-Erik Thorelli, Sweden).*

Regular Subscriptions:

The annual institutional subscription price for Volume 5, 1979, published in 6 issues, is US $61.00/Dfl. 125.00.

EUROMICRO JOURNAL is published in bi-monthly issues, appearing in January, March, May, July, September and November.

FREE specimen copies are available on request from the publisher.

north-holland

P.O. BOX 211 / 1000 AE AMSTERDAM / THE NETHERLANDS

IN THE U.S.A. AND CANA / 52 VANDERBILT AVENUE / NEW YORK, N.Y. 10017

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

Analytica Chimica Acta, 112 (1979) 199—209
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

QUALITATIVE CLASSIFICATION OF DITHIOCARBAMATE COMPOUNDS FROM ^{13}C-N.M.R. AND I.R. SPECTROSCOPIC DATA BY PATTERN RECOGNITION TECHNIQUES

F. W. PIJPERS*

Department of Analytical Chemistry, Catholic University, Toernooiveld, Nijmegen (The Netherlands)

H. L. M. VAN GAAL and J. G. M. VAN DER LINDEN

Department of Inorganic Chemistry, Catholic University, Toernooiveld, Nijmegen (The Netherlands)

(Received 13th February 1979)

SUMMARY

Unsupervised pattern recognition permits qualitative classification of *N,N*-dialkyldithiocarbamate complexes into five classes, denoted as free ligands, complexes of main group elements, and complexes of transition metals both with FON = 0.5 or FON > 0.5. Within the transition metal complexes, metals with electron configuration d^6 or d^8 and d^0 or d^{10} can be further distinguished. The class of main group elements is subdivided into organic dithiocarbamates including PE_3 and TeE_2 (E = Et_2DTC) and a subclass of main group element complexes. Supervised techniques based on these five classes show that the descriptors $\delta(N^{13}CS_2)$, $\nu(C{=}N)_1$ and atom number are dominant in classifying compounds. The substituent descriptor and $\delta(\alpha\text{-}^{13}CH_2)$ are less important. The classification and ranking of the descriptor quality gives an insight into the relationship between the quantities measured. Classifications which disagree with chemical expectations are readily identified and this improves understanding of the chemical background.

A detailed study concerning the factors that influence the ^{13}C-n.m.r. chemical shifts, $\delta(N^{13}CS_2)$, of the NCS_2-fragment of 65 diamagnetic alkyldithiocarbamates (DTCs) has recently been reported [1]. This study showed that the major part of these shifts is determined by the π-bonding in the NCS_2 moiety, and that the coordination mode, symmetrical or otherwise, of the coordinated DTC definitely affects the $\delta(N^{13}CS_2)$. The bonding mode is also reflected in the magnitude of the carbon—nitrogen stretching vibration frequency, $\nu(C{=}N)$, for the large group of compounds studied. Among other factors to which attention was paid, were the electronic configuration, the oxidation number and the coordination number of the coordinated element. All these quantities seem to be mutually related. The apparent complexity of these relations suggested the use of pattern recognition techniques for their detection. For this purpose, a computer program ARTHUR [2] was applied to the $\delta(N^{13}CS_2)$ values, $\nu(C{=}N)$ values and several other descriptors of 65 compounds, in order to arrive at a classification of the compounds. The pat-

ห้องสมุด กรมวิทยาศาสตร์บริการ
17. ธ.ค. 2522

tern recognition program defines the relative importance of the various descriptors for classification, gives an indication whether the classes are properly defined, and gives a numerical correlation between descriptors. Furthermore, it offers the possibility of classifying unknown and unclassified compounds, permitting the detection of anomalies in the unknown compounds. The procedures applied are described in this paper.

EXPERIMENTAL

The data base was taken from a previous paper [1]. The pattern recognition techniques applied were: (i) Zahn minimal spanning tree construction (TREE) [3] and hierarchical or Q-mode clustering (HIER) [4, 5] for unsupervised learning; and (ii) Bayes theorem classification (BAYES), nearest neighbor classification (KNN) [4] and linear pattern classification (PLANE) [6—8] for supervised learning. The features of the data base were preprocessed by CORREL [9] which is a method for detection of inter-feature correlation, SCALE [4, 5] which is a method for the production of unbiased features, WEIGHT [4, 5, 10] which deals with feature ranking, and KARLOV [11] which calculates eigenvalues and eigenfunctions based on feature variances. The patterns were placed in a multidimensional representation space by DIST [12] and projected by means of the non-linear mapping procedure NLM [4, 5, 11] with the plotting routine VARVAR. These methods of calculation are all part of the ARTHUR program [2] and are denoted by the names used there.

RESULTS AND DISCUSSION

The classes, with unsupervised clustering

In order to define classes, the HIER and TREE methods were applied in an unsupervised fashion on a training set of 56 compounds. The other nine compounds of the data set formed an evaluation set, selected for chemical reasons (see below). Similarity values, $S_{i,j}$, based on their mutual distance, $D_{i,j}$, are defined as: $S_{i,j} = 1 - D_{i,j}/D_{max}$. The $D_{i,j}$ are Euclidean distances in the five-dimensional space spanned by the features $\delta(N^{13}CS_2)$, $\nu(C\overset{\cdots}{=}N)_{liquid}$, $\delta(\alpha\text{-}^{13}CH_2)$, substituent number and atom number. The numerical values of the latter two features are explained later. The compounds are grouped at levels of similarity, $S_{i,j}$, in HIER and are linked in branched chains based on the distances, $D_{i,j}$, in TREE. Clusters related to the free ligands, to the transition metal complexes, and to compounds of the main group elements could be detected, but a much more detailed classification was also obtained. In order to achieve a chemical description of the properties of the calculated classes, apart from the characteristics already mentioned, the fractional oxidation number, FON, was introduced as a second class denominator. FON is defined as the ratio between the oxidation number and coordination number of the central metal. As such, it turned out to be a good way of introducing

the influence of coordination number as well as oxidation number, which separately could not be related to the class divisions. FON proved to be a promising concept, as has already been extensively discussed [1]. Classes were constructed with FON = 0.5 and with FON > 0.5 of compounds of main group elements on the one hand and of transition metal complexes on the other hand, resulting in four classes. Together with the free ligands, this resulted in five different classes which each coincided with one of the major clusters calculated by HIER and by TREE, apart from a few exceptions. These classes are summarized in Table 1, together with the ranges of values of the three most important descriptors.

The classes thus found were used for modification and selection of the descriptors prior to a more refined supervised application of the HIER and TREE methods.

The descriptors

Five features were used as descriptors: three measured quantities, i.e. $\nu(C{\cdots}N)_{liquid}$, $\delta(N^{13}CS_2)$ and $\delta(\alpha\text{-}^{13}CH_2)$ for which the numerical values have been presented elsewhere [1], a substituent number and an atom number. The substituent numbers are: 1 = ethyl; 2 = n-butyl and 3 = other substituents. The atom numbers are 4, 5 or 6 corresponding to the period number of those central metals having electron configuration d^6, d^8 or $d^{10}s^2$, and 3 for all other complexes.

All descriptors $x_{i,j}$ (feature i of compound j) were autoscaled to a mean of zero and a normalized standard deviation according to

$$x'_{i,j} = (x_{i,j} - \bar{x}_i)/\sigma_i(N-1)^{\frac{1}{2}}$$

where $x'_{i,j}$ is the new, autoscaled descriptor of feature i for compound j, $\bar{x}_i$ is the average value of descriptor i, σ_i is the standard deviation of the ith descriptor

TABLE 1

The classes of the training set

Sample class and a posteriori description	FON	N_{class}	Feature boundaries		
			$\nu(C{\cdots}N)_{liquid}$ (cm^{-1})	$\delta(N^{13}CS_2)$ (ppm)	Atom no.
Free ligands					
Class 1	—	4	1466—1485	207—212	3
Transition metal DTCs					
Class 2: normal valence states	0.5	25	1480—1520	202—212	3,4,5,6
2a: d^6 and d^8 complexes		(15)	1490—1520	205—212	4,5,6
2b: d^0 and d^{10} complexes		(10)	1480—1501	202—206	3
Class 3: high valence states	>0.5	7	1520—1555	193—201	4,6
Main group and organic DTCs					
Class 4: normal coordination numbers	0.5	13	1480—1496	193—203	3,4,5,6
4a: main group compounds		(8)	1480—1494	197—203	4,5,6
4b: organic DTCs, PE_3, TeE_2[a]		(5)	1483—1496	193—195	3,5
Class 5: low coordination numbers	>0.5	7	1499—1529	192—196	4,5

[a]E = Et_2DTC.

and N is the number of compounds in the training set. The correlation between various features, cor(i, j) was calculated from

$$\mathrm{cor}(i,j) = \sum_{k=1}^{N} (x_{i,k} - \bar{x})(x_{j,k} - \bar{x}_j) \Big/ \left[\left\{ \sum_{k=1}^{N} (x_{i,k} - \bar{x}_i)^2 \right\} \left\{ \sum_{k=1}^{N} (x_{j,k} - \bar{x}_j)^2 \right\} \right]^{\frac{1}{2}}$$

In addition to the correlation between the five descriptors mentioned, the correlation with two other features was calculated: $\nu(\mathrm{C{\cdots}N})_{\mathrm{solid}}$ and FON, which was temporarily used as a feature apart from its use as a class denominator. The correlation coefficients thus obtained are listed in Table 2, which shows that the FON is highly correlated with $\nu(\mathrm{C{\cdots}N})_{\mathrm{liquid}}$, $\nu(\mathrm{C{\cdots}N})_{\mathrm{solid}}$ and with $\delta(\mathrm{N}^{13}\mathrm{CS}_2)$. This means that a valid independent classification can be made without FON. Another high correlation is that between $\nu(\mathrm{C{\cdots}N})_{\mathrm{liquid}}$ and $\nu(\mathrm{C{\cdots}N})_{\mathrm{solid}}$. This makes one of these two features superfluous; $\nu(\mathrm{C{\cdots}N})_{\mathrm{solid}}$ was therefore eliminated.

Feature weighting

Based on the mean values and standard deviations within each class, the autoscaled descriptors were weighted with variance and Fisher weighting factors, to establish their relative importance for classification. The Fisher weighting factor $W(F)_{j,m,n}$ for separating class m from class n by means of descriptor j is

$$W(F)_{j,m,n} = [\overline{(x'_m)_j} - \overline{(x'_n)_j}]^2 / (N_m \sigma^2_{m,j} + N_n \sigma^2_{n,j})$$

in which $\overline{(x'_i)_j}$ denotes the mean value of the autoscaled descriptor j in class i, N_i is the number of complexes in class i, and $\sigma_{i,j}$ is the standard deviation of the autoscaled descriptor j in class i. The Fisher weighting factor of descriptor j for all linear class separations, $W(F)_j$, is

$$W(F)_j = 2 \left\{ \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} W(F)_{j,m,n} \right\} \Big/ M(M-1)$$

in which M denotes the number of classes ($M = 5$).

The variance weighting factor, $W(V)_{j,m,n}$, for separating class m from class n by means of descriptor j is

$$W(V)_{j,m,n} = [\overline{(x'_m)_j^2} + \overline{(x'_n)_j^2} - \overline{2(x'_m)_j}\ \overline{(x'_n)_j}] / [(m2)_{m,j} + (m2)_{n,j}]$$

TABLE 2

Intercorrelation between descriptors

	$\delta(\mathrm{N}^{13}\mathrm{CS}_2)$	$\nu(\mathrm{C{\cdots}N)_1$	Atom no.	$\delta(\alpha\text{-}^{13}\mathrm{CH}_2)$	Subst. no.	FON	$\nu(\mathrm{C{\cdots}N})_{\mathrm{s}}$
$\delta(\mathrm{N}^{13}\mathrm{CS}_2)$	—	−0.31	−0.09	−0.15	0.38	−0.60	−0.23
$\nu(\mathrm{C{\cdots}N})_{\mathrm{liquid}}$		—	0.51	−0.18	0.08	0.62	0.88
Atom number			—	0.12	0.09	0.16	0.35
$\delta(\alpha\text{-}^{13}\mathrm{CH}_2)$				—	0.44	0.10	−0.22
Substituent no.					—	−0.01	0.09
FON						—	0.66
$\nu(\mathrm{C{\cdots}N})_{\mathrm{solid}}$							—

in which $(m2)_{i,j}$ is the second moment of descriptor j in class i: $(m2)_{i,j} = (N_i - 1)\sigma^2_{i,j}/N_i$. The variance weighting of descriptor j for all linear class separations, $W(V)_j$, is

$$W(V)_j = \left[\prod_{m=1}^{M-1} \prod_{n=m+1}^{M} W(V)_{j,m,n} \right]^{2/M(M-1)}$$

The descriptor weightings are listed in Table 3. It is clear from both the variance weighting and the Fisher weighting that $\delta(N^{13}CS_2)$, $\nu(C{\equiv}N)_1$ and the atom number are the three most important features for classification.

The training set

After the features had been selected and their relative importance for classification determined, the classifications based on the hierarchical clustering mode and on the minimal spanning tree were repeated, this time in a supervised calculation, by using the five descriptors variance weighted for the five classes. The clusters thus obtained were essentially the same as in the unsupervised calculation. They are presented in Table 4 for HIER and in Fig. 1 for TREE (see also Table 5). In HIER, class 2 is subdivided into two smaller classes, one with high mutual similarity level (0.88) and the other with compounds of class 2 at a similarity level of 0.66. Class 4 is also separated into two sub-classes, one with a similarity level of 0.86 and the other with a similarity level of 0.88, both combined at a level of 0.73. Class 3 gains two compounds out of class 5: $Sn(Et_2DTC)_3BPh_4$ (compound 50 from Table 5) and $As(n\text{-}Bu_2DTC)Br_2$ (compound 52).

In TREE, all compounds are ordered according to the five classes; the subclasses 2a, b and 4a, b are also present. Figure 1 shows the spanning tree, in which the length of the interconnecting lines between the numbers represent distances. The numbers refer to the compounds given in Table 5. As well as the ordering presented, TREE gives three clusters in the supervised method. The clustering combines all class 2 compounds together with one class 1 compound (number 2) into one cluster. The latter compound, $NaEt_2DTC \cdot 3H_2O$, is positioned at a distance 0.6823 from the nearest class 1 compound and at a distance 0.6852 from the nearest class 2 compound, but the larger

TABLE 3

Fisher and variance weightings, and eigenfunctions based on the variance weightings for selected descriptors

Descriptor	$W(F)_j \times 10^2$	$W(V)_j$	% of total	Eigenvectors of the $W(V)_{j,n,m}$ matrix (columns)				
$\nu(C{\equiv}N)_{liquid}$	92.1	6.73	37.3	0.83	0.51	−0.18	−0.11	0.09
$\delta(N^{13}CS_2)$	122.1	5.86	32.6	−0.54	−0.83	−0.05	−0.05	0.14
Atom no.	28.2	2.40	13.3	0.14	−0.18	0.81	0.54	−0.01
Substituent no.	4.7	1.52	8.4	−0.01	−0.12	−0.40	0.56	−0.71
$\delta(\alpha\text{-}^{13}CH_2)$	8.6	1.51	8.4	−0.01	0.10	−0.38	0.62	0.68
Eigenvalues, $Info(E)_k$, (% of total) of $W(V)_{j,n,m}$ matrix				62.2	29.1	5.1	2.9	0.7

TABLE 4

Hierarchical clustering of compounds of the training set, based on variance-weighted descriptors

Class	N_{class}	Similarity values
2a	10	0.88
2b-1	5	0.89
		0.82
4a	8	0.86
		0.79
4b	5	0.88
		0.82
5-1	5	0.91
		0.73
2b-2	10	0.78
		0.66
1	4	0.86
		0.64
		0.51
3-1	2	0.86
		0.79
5-2	2	0.91
		0.72
3-2	5	0.87

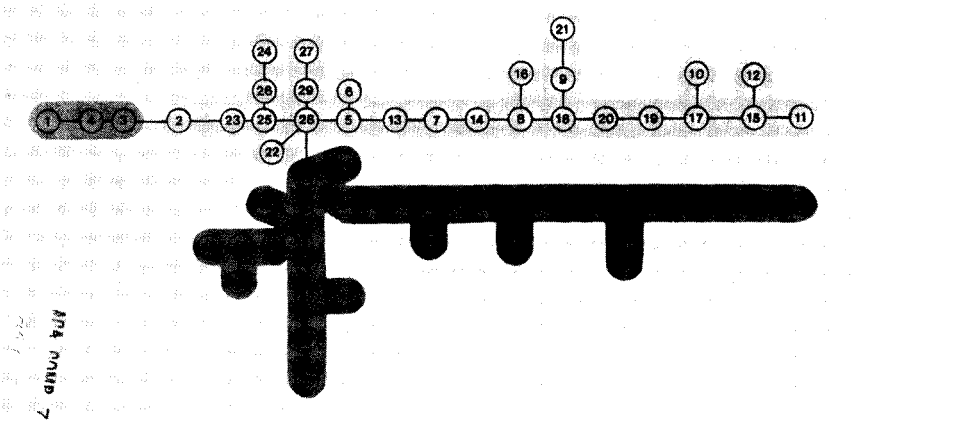


Fig. 1. Zahn minimal spanning tree of training set patterns based on Euclidean distances and covariance-weighted descriptors.

population of class 2 tips the balance of decision towards this group in clustering compound 2 of class 1 with the members of class 2. The remaining three compounds of class 1 form a second cluster, and classes 3, 4 and 5 form the third cluster.

TABLE 5

Compounds of the training set. Numbers refer to Fig. 1

Number	Compound[a]	Class	Number	Compound	Class
1	n-Bu_4NE	1	35	$AuLCl_2$	3
4	n-Bu_4NL	1	34	$AuLBr_2$	3
3	n-Bu_2NH_2L	1	36	AuL(MNT)	3
2	$NaE.3H_2O$	1	33	AuL_2AuBr_2	3
23	$(CuL)_4$	2a	31	PtE_3ClO_4	3
25	$(AgL)_6$	2a	32	PtL_2I_2	3
26	$(AuL)_2$	2a	30	NiL_3ClO_4	3
24	$[Ag(n\text{-}Pr_2DTC)]_6$	2a	53	AsL_2I_2	5
28	CdL_2	2a	50	SnE_3BPh_4	5
29	HgL_2	2a	52	$AsLBr_2$	5
27	ZnL_2	2a	55	$SbLBr_2$	5
22	$(CuE)_4$	2a	56	SbL_2I_3	5
5	$Ti(n\text{-}Pr_2DTC)_4$	2a	51	AsL_2Br	5
6	$Zr(n\text{-}Pr_2DTC)_4$	2a	54	SbL_2Br	5
13	$Ni(i\text{-}Pr_2DTC)_2$	2b	42	AsL_3	4a
7	CoL_3	2b	44	SbL_3	4a
14	NiL_2	2b	41	AsE_3	4a
8	RhL_3	2b	46	TeE_2	4b
16	$Ni(Bz_2DTC)_2$	2b	48	$(ECH_2)_2$	4b
18	PdL_2	2b	49	LCH_2COPh	4b
9	IrL_3	2b	47	*N,N,N',N'*-tetraethylthiuram disulfide	4b
21	$Pt(i\text{-}Bu_2DTC)_2$	2b			
20	PtL_2	2b			
19	PtE_2	2b	40	PE_3	4b
17	PdE_2	2b	43	SbE_3	4a
10	NiE_2	2b	38	SnE_2	4a
15	$Ni(n\text{-}BuHDTC)_2$	2b	45	BiE_3	4a
12	$Ni(n\text{-}PrHDTC)_2$	2b	37	$(TlE)_2$	4a
11	$Ni(n\text{-}Pr_2DTC)_2$	2b	39	PbL_2	4a

[a]E = Et_2DTC; L = n-Bu_2DTC.

Two-dimensional mapping

In order to visualize the classes and the intra-classes pattern distribution, various display methods are available in the ARTHUR program. An attractive method is offered by the Karhunen—Loève transformation KARLOV [9], which constructs eigenvectors as linear combinations of descriptors, on the basis of the descriptor weight matrices. The variance weight matrix, $W(V)_{j,m,n}$, was applied for this transformation. The percentage amounts of information contained in the kth eigenvector, Info $(E)_k$ and the eigenvectors are listed in Table 3. The eigenvectors are ordered by decreasing variance eigenvalues. The first two eigenvectors yield 91.3% of all information. This is an improvement of 21.4% compared to the information content of $\nu(C{\cdots}N)$ and $\delta(N^{13}CS_2)$ which is 69.9%. Figure 2 presents the projection of the data of training and evaluation set on a plane defined by the first two eigenvectors of $W(V)_{j,m,n}$. Visual inspection of this projection suffices for classification.

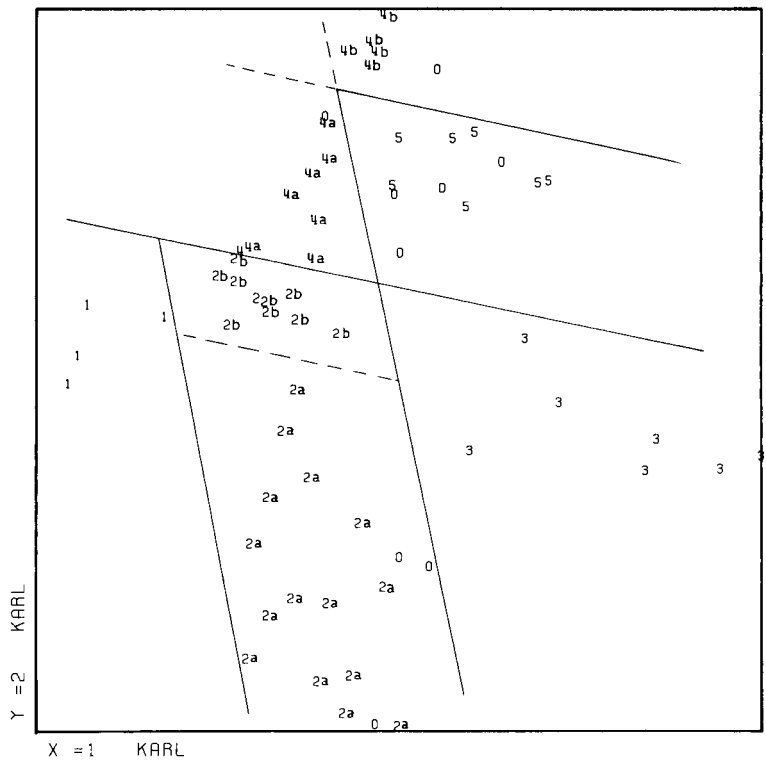


Fig. 2. Projection of the training and evaluation set patterns on a plane defined by the most important eigenvectors of the Karhunen—Loève transformation. Classes are depicted by full and dashed lines. The zero points denote compounds from the evaluation set.

The non-linear mapping of the training set on a plane (Fig. 3) by NLM [4, 5, 10] results in a display that resembles the KARLOV projection of Fig. 2 (after a 180° rotation). Here too the boundaries between classes are drawn after visual inspection. The results confirm the classes given by TREE and somewhat improve those given by HIER, especially with respect to the separation of classes 3 and 5.

Chemical significance of the classes

The oxidation and coordination numbers are described in Table 1 by "normal", "high" and "low", which can be related to the FON values. "Normal" corresponds to a FON value of 0.5. A high FON value refers to a high oxidation state or a low coordination number, and indicates a high positive charge on the central metal atom. Thus the FON value provides a means of distinguishing class 2 from class 3 and class 4a from class 5 (see, e.g., Fig. 3). It is a merit of the pattern recognition calculations that the group of hitherto-unnoticed "low coordination number main-group DTCs" has been revealed

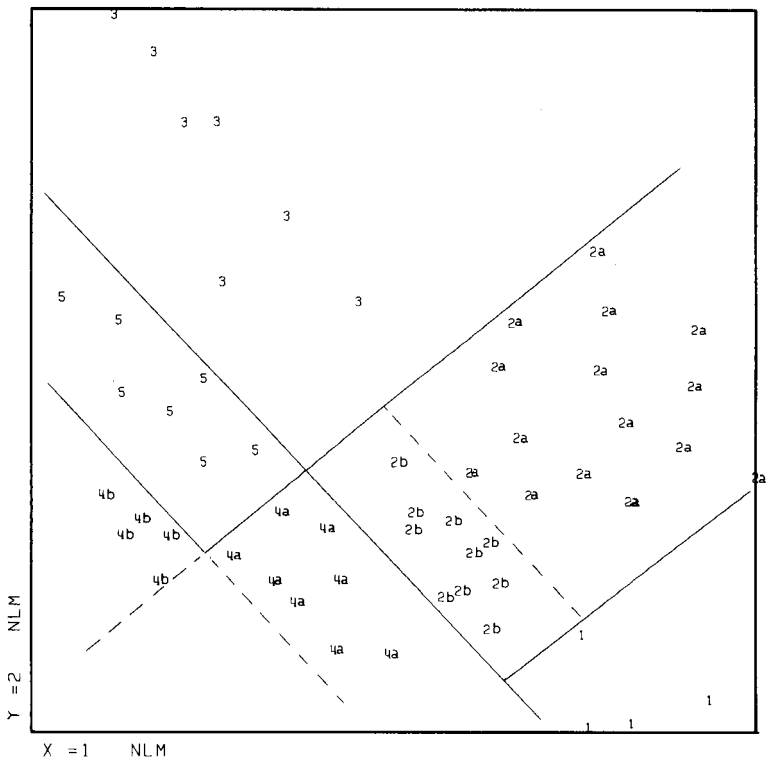


Fig. 3. Non-linear mapping of the training set patterns from a five-dimensional space with Euclidean distances.

and that this group has been associated with the "high oxidation state transition metal DTCs". Structural data show that, rather than the difference between transition metals and main group elements, the difference between symmetrical and asymmetrical dithiocarbamate bonding is the fundamental property distinguishing class 2 from class 4 and class 3 from class 5. Moreover, the difference in the degree of asymmetry of the dithiocarbamate causes a subdivision of class 4 into 4a (main group elements) and 4b (organic DTCs). The subdivision of the class 2 compounds is of a different nature: all 2a complexes have d^0 and d^{10} metal electronic configurations, whereas 2b compounds have d^6 and d^8 configurations. The differences in excitation energy associated with the various configurations, affect the $\delta(N^{13}CS_2)$ values differently for class 2a and class 2b complexes.

Other supervised classification methods

The training set. With the established classes 1—5, supervised classification methods were applied. One of those methods operates according to the Bayes rule. The Bayes theorem probability that a given compound in a member of class k based on descriptor j and its associated probability distribution is given

$$P_j[X_{j,k}|x_{i,j}] = P[x_{i,j}|X_{j,k}] \Big/ \sum_{n=1}^{M} P[x_{i,j}|X_{j,n}]$$

where $P[x_{i,j}|X_{j,n}]$ represents the value of the probability distribution $X_{j,n}$ for descriptor j of class n at the value $x_{i,j}$. The probability distributions follow from the frequency histograms of the compounds in the data set. In this investigation, the probability of three descriptors $\nu(C{\equiv}N)$, $\delta(N^{13}CS_2)$ and atom number, were combined according to

$$P[X_k|x_i] = \{1/A(\alpha)\} \sum_{j=1}^{3} \{P_j[X_{j,k}|x_{i,j}]\}^{\alpha} \qquad (\alpha = \tfrac{1}{2}, 1, 2)$$

where $A(\alpha)$ represents a normalization constant. The square root was proved empirically to give the best class separation.

This means that complexes situated near the boundaries of the probability distribution curves are considered with enhanced probability to belong to that particular distribution: the probability distribution curve is broadened by taking the square root. Compounds of classes 1, 4 and 5 were completely separated from the other classes. Class 2 missed four compounds, three of which were placed in class 4 [$Rh(n\text{-}Bu_2DTC)_3$, $Cd(n\text{-}Bu_2\text{-}DTC)_2$ and $Hg(n\text{-}Bu_2DTC)_2$] and one of which [$Pd(n\text{-}Bu_2\text{-}DTC)_2$] was classified in class 5. Class 3 missed one compound [$Ni(n\text{-}Bu_2DTC)_3ClO_4$] which according to the Bayes classification was placed in class 5. Although these results are in reasonable agreement with those obtained by the other techniques applied, it remains questionable whether this statistical technique is really validated with such a limited number of patterns per class. Excellent results by this method have been reported in the literature with more patterns per class (about 250—550) [13].

Another supervised classification of the training set was based on the nearest neighbor of a compound vector in the n-dimensional space by using distances $D_{i,j}$ as previously defined. The results of 1—5 nearest neighbor classification are considered here. The compounds of classes 2, 4 and 5 were unanimously and correctly classified by their neighbors. Because class 1 contained only four patterns, one compound [$NaEt_2DTC \cdot 3H_2O$] was twic misclassified in class 2. Class 3 always lost one compound [$Ni(n\text{-}Bu_2DTC)_3ClO_4$] to class 5.

A third classification with supervision was based on a binary classifier PLANE, that separates all possible pairs of classes (here ten) by using negative feedback training. The classification with PLANE indicates the quality of the training set, because when one or more class separations are unsuccessful, the quality of the training set classification is usually poor, and its success as a classifier is limited. The separation with PLANE was 100% successful for all pairs of classes.

The evaluation set. The descriptors that were developed with the training set were applied for the classification of an evaluation set of nine complexes that were peculiar from a chemical point of view. The classifications are listed in Table 6. The position of $VO(Et_2DCT)_3$ indicates that the compound behaves anomalously; the position of $(PPh_3)Au(Et_2DTC)$ is such that this

TABLE 6

Classification of compounds of the evaluation set by various methods

Compound[a] E = Et_2DTC	PLANE	KARLOV projection 1-2	KNN	BAYES class (probability, %)
VOE_3	5	5	5	5 (0.32)
$(PPh_3)AuE$	2	2	2	1 (0.31) 2 (0.27)
SnE_4	5	5	2 (5)	2 (0.26) 5 (0.25)
$HgI_2(Et_4TDS)$	5	5	5	5 (0.52)
$MeAsE_2$	4	4	4	5 (0.28) 4 (0.24)
$Ir(C_8H_{12})E$	2	2	2	3 (0.34) 2 (0.27)
$Ir(C_8H_{12})E_3$[a]; b	2	2	2 (3)	3 (0.34) 2 (0.31)
$Ir(C_8H_{12})E_3$[a]; m_1	4	5	4	3 (0.37) 4 (0.28)
$Ir(C_8H_{12})E_3$[a]; m_2	4	5	4	4 (0.34)

[a]E = Et_2DTC. b is bidentate; m_1 and m_2 are non-equivalent monodentate DTC ligands.

compound should have a solution structure different from the solid-state structure; $Sn(Et_2DTC)_4$, a fluxional molecule, is not readily classifiable; and $HgI_2(Et_4TDS)$, the only *N,N,N',N'*-tetraethylthiuram disulfide complex, which before classification seemed similar to organic dithiocarbamates, had to be reclassified as being similar to the main-group dithiocarbamates of low coordination number. The position of the compounds $AsMe(Et_2DTC)_2$, $Ir(Et_2DTC)(C_8H_{12})$ and $Ir(m\text{-}Et_2DTC)_2$ $(b\text{-}Et_2DTC)$ (C_8H_{12}) (m = monodentate, b = bidentate bonded DTC) could be interpreted on the basis of the electronic properties of the co-ligands and of the bonding mode of the dithiocarbamate ligand. On this basis, an improved molecular orbital description of the effect of π-bonding in the NCS_2 fragment on the $\delta(N^{13}CS_2)$ value proved feasible [1].

We thank Prof. G. Kateman, Prof. Dr. Ir. J. J. Steggerda and Ir. B. G. M. Vandeginste for their critical remarks, and Prof. B. R. Kowalski for making the ARTHUR program available.

REFERENCES

1 H. L. M. van Gaal, J. W. Diesveld, F. W. Pijpers and J. G. M. van der Linden, Inorg. Chem., accepted for publication.
2 D. L. Duewer, J. R. Koskinen and B. R. Kowalski: Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, Washington 98195.
3 H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Wiley—Interscience, New York, 1972.
4 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.
5 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
6 N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965; D. R. Preuss and P. C. Jurs, Anal. Chem., 46 (1974) 520.
7 T. L. Isenhour, B. R. Kowalski and P. C. Jurs, C.R.C. Crit. Rev. Anal. Chem., 4 (1974) 1.
8 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 690
9 O. L. Davies and P. L. Goldsmith, Statistical Methods for Research and Production, Hafner, 1972, p. 234.
10 R. A. Fisher, Ann. Eugen., 7 (1936) 179.
11 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.
12 T. L. Isenhour and P. C. Jurs, Anal. Chem., 43 (1971) 20A.
13 C. L. Wilkins and T. R. Brunner, Anal. Chem., 49 (1977) 2136.

Analytica Chimica Acta, 112 (1979) 211—218
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

AN ALGORITHM FOR ASTM INFRARED FILE SEARCHES BASED ON INTENSITY DATA

K. TANABE*, T. TAMURA, J. HIRAISHI and S. SAËKI
National Chemical Laboratory for Industry, Tsukuba, Ibaraki 300-2 (Japan)

(Received 20th March 1979)

SUMMARY

The computer program described utilizes peak intensity data for ASTM infrared file searching. Peaks in unknown spectra are classified into five groups according to their relative intensities, and the scores for matches with the ASTM data are calculated by means of the intensity data. A test search with 135 compounds proves the excellent performance of the proposed method. The advantages of the system are that correct answers can be found easily and that no particular attention is necessary for selection of the peaks to be entered.

Numerous computer search systems have been proposed for infrared spectral data. In one type, which includes pattern recognition techniques [1—9], and the correlation coefficient methods [10—13], digitalized data of the spectral patterns as a whole are utilized. In the second type of system, some indexed data extracted from the original spectral pattern are utilized in the matches; examples are the ASTM codes [14, 15] and the techniques based on centers of gravity and standard deviations in spectra [16], or on peak position and intensity data [17—19].

Computer search systems based on the ASTM data file are now widely used commercially because the file includes spectral data for more than 100,000 compounds. Numerous systems [20—40] have been reported for ASTM file searching, but the intensity data in unknown spectra do not seem to be utilized effectively for matching in the systems proposed so far. Although the ASTM data do not contain any intensity information (except for masking if a group of peaks includes one or more strong peaks), the use of intensity data in the matches will improve the efficiency of a search process.

In this paper, a new computer program is reported which utilizes data on peak intensity as well as peak position for matches with the ASTM file. A test search was done for more than 100 known compounds as a sample, to assess the efficiency of the system.

METHODS

All peaks in unknown spectra are classified into five groups according to

their relative intensities. Class 5 denotes the strongest peaks, and class 1 the weakest. The absorbance scale is used for the measurement of peak intensities as in the ASTM data. The input data for the class 5 peaks and of the no-peak areas are treated as a mandatory condition, and the other data are used as a desirable condition, in the matching procedures.

The score for a hit spectrum in the file is computed from the equation

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

where N_i is the number of hit peaks in the i-th intensity class, and p is defined as a penalty number which denotes the highest intensity class where the number of hit peaks is smaller than that of input peaks.

The principle of eqn. (1) is as follows. When a peak of some particular intensity in the unknown spectrum is missing in one of the file spectra, hit peaks which have lower intensities have to be regarded as accidentally or erroneously hit. Therefore those hits should be given a penalty number rather than a score. Equation (1) is based on such a scoring procedure, and its use is effective in reducing noise and improving the performance of search.

Examples of the scoring are illustrated in Fig. 1. The spectral pattern of the correct compound 1 completely agrees with that of the unknown sample, so that a score of 100 points is given. When a reference spectrum 2 of the correct compound is measured more weakly than the unknown, 100 points are scored. In contrast, noise always receives a lower score, even if the number of hit peaks is greater than that in the correct compound. In fact, the noise pattern 3 scores quite low because of the presence of hit peaks below $p = 4$:

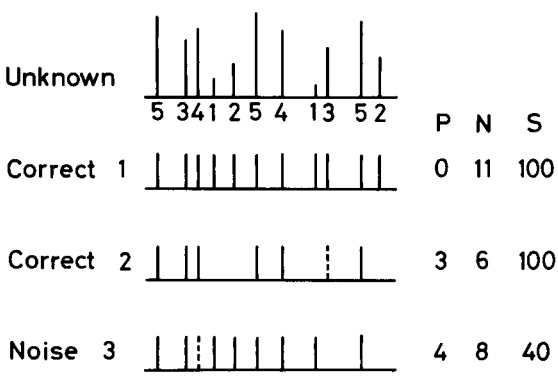


Fig. 1. Illustration of the scoring method of eqn. (1). The top spectrum shows the relative peak intensities of the unknown sample together with their intensity classes. The next three show the spectral patterns of the three hits in the ASTM file. Solid lines denote "on" bits in the file data (punched on the card base), and broken lines denote bits which correspond to the strongest peaks in the unknown but are not punched in the ASTM cards. P is the penalty number (see text), N is the total number of hit peaks, and S is the score computed by means of eqn. (1). It is assumed that the spectrum 1 of the correct compound is measured under experimental conditions similar to those for the unknown sample, and that spectrum 2 is observed more weakly. Spectrum 3 represents noise satisfying the mandatory conditions.

$$S = 100\ \{5 \times 3 - [(4-3) \times 1 + (4-2) \times 1 + (4-1) \times 2]\}/\{5 \times 3\} = 40 \text{ points}$$

Input data

The input data in the present system are peak frequencies with their classes as mentioned above, and frequency ranges of no-peak areas. The peak position data are entered as the wavenumbers of the maximum absorptions, and in the program, the appropriate tolerance (see below) is added for the matching procedures. The input data of the no-peak areas are entered as the wavenumbers of the two peaks lying on either side of the areas, and in the program, the areas are diminished by the wavelength tolerance of both peaks. If the input no-peak areas vanish after the procedure, the data are automatically excluded from the input list.

In order to determine the frequency tolerance experimentally, the frequency shifts were examined by comparing the ASTM data with the observed spectra. It was found that the curve shown in Fig. 2. covers peak shifts encountered in the searches for most of the unknown samples. On this basis, in the program, the peak position data are converted to wavelengths and the tolerance is added according to the equation

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

where λ_{max} is the peak wavelength in μm. This equation reproduces the tolerance curve shown in Fig. 2 quite well.

Test search

Infrared spectra of 135 compounds were measured with a Shimadzu IR-27G grating spectrophotometer. The compounds were selected from chemicals to hand, with attention to impartiality in chemical classes. Spectra of 80 liquid

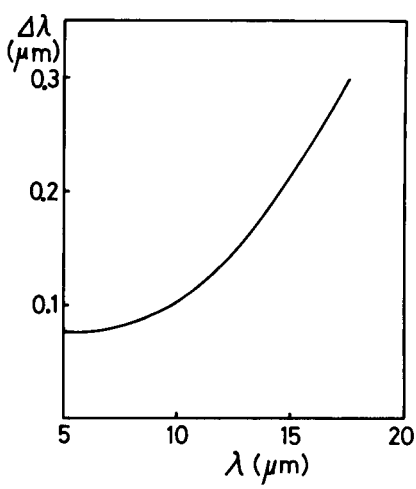


Fig. 2. Tolerance curve for peak shifts in wavelength.

compounds were obtained from capillary films, and those of 55 solid compounds from KBr disks.

The search program was written in FORTRAN IV, except for part of the bit comparison written in assembler language. The size of the total program is about 15 k words. A Yokogawa-Hewlett-Packard model 2112A computer with 65 k word main memory and a 5 megabyte disk memory was used.

Spectral data files of more than 100,000 compounds are now published by the ASTM. In this study, about 55,000 data in hand (Table 1) were used for the test search. Expansion of the file data to more than 100,000 spectra will not change the results obtained for the search.

RESULTS AND DISCUSSION

An example of the test search is summarized in Table 2, and the observed spectrum of the sample (indole) is shown in Fig. 3. The correct spectra (545FA, 18034CA, and 3229JA) obtain 100 points, which allows them to be retrieved and identified as the correct answers. The test search for 135 known samples showed that the correct spectra always receive the full score. This indicates the advantage of the proposed system: correct answers will always be found among spectra of the highest scores.

In many earlier search systems, it seems that some experience is required for satisfactory results as indicated by Erley [28]. In these systems, the efficiency of search depends on the selection of peaks to be entered. If the number of input peaks is much smaller than the number necessary for a successful search, numerous noise spectra will be listed and the correct answers will become obscured.

Conversely, if too many peaks are entered, the correct answers will be ranked too low to be found easily. Therefore, the number of input peaks for successful search in these systems must be estimated according to the criteria of the ASTM coding.

For the ASTM coding, all peaks showing an absorbance ratio of 1:10 or more to the absorbance of the strongest peak in the spectrum should be

TABLE 1

ASTM data file used in the test search

File name	Mark	Number of spectra
Sadtler	CA	25 000
DMS	FA	9 900
IRDC	JA	5 000
Coblentz	GA	3 000
API	AA	2 600
NBS	DA	2 500
Others		7 000
Total		55 000

TABLE 2

An example of the test search for a sample of indole (see Fig. 3)

Input data Peaks (number = 23)				*Output*		
Wave-number	Class	Wave-number	Class	Rank	Serial number	Score
1704	1	1150	1	1	18034CA[a]	100
1612	1	1120	2	2	3229JA[a]	100
1573	1	1091	4	3	545FA[a]	100
1501	1	1060	4	4	2819AA[a]	94
1485	1	1004	2	5	2820AA[a]	93
1452	5	934	3	6	6425CA	93
1412	5	897	3	7	358FA	92
1352	5	862	2	8	550DA	92
1337	5	765	3	9	855GA	92
1278	3	746	5	10	1374GA	90
1247	4	723	5	11	3935FA	88
1204	2			12	14267CA	85
				13	10823CA	83
No-peak areas (Number = 5)				14	12040CA	83
				15	16693CA	83
Wavenumber range				16	9563CA	80
				17	821JA	80
1573—1501				18	625FA	80
1204—1150				19	1062DA	80
1004—934				20	2716DA	80
862—765						
723—650						

[a]Correct answers.

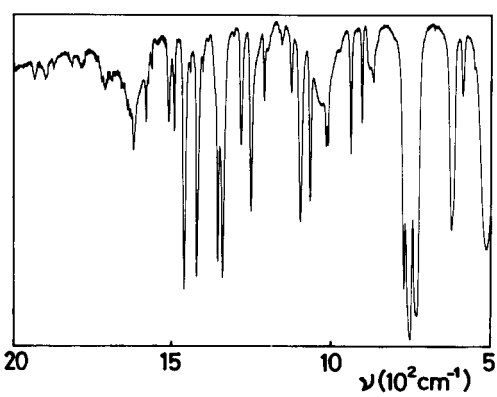


Fig. 3. Infrared spectrum of indole for a test search (KBr disk).

ห้องสมุด กรมวิทยาศาสตร์บริการ

punched in the ASTM cards [15]. When many of the ASTM data were compared with the present original spectra, however, the criteria were not always fulfilled in the coding. Therefore, when these search methods are applied, much attention must be given to entering an appropriate number of data without listing weak peaks.

Such attention is unnecessary for the proposed system, and all peaks, even those of very low intensity, can be added to the input list. This is another advantage of the proposed search system. In terms of better search results, it is desirable to enter as many peaks as possible. When the distribution of the numbers of peaks punched in the ASTM cards was examined, it was found (Fig. 4) that it is sufficient to enter 20—25 peaks as the input data.

Since the input data for the no-peak areas are treated as mandatory conditions in the matching procedures, it is better to enter them with as much restraint as possible. The wavelength scale is used to indicate peak positions in the ASTM data, and the matches in the present system are also based on the wavelength scale. Therefore, when two or more no-peak areas have approximately equal lengths of frequency range, it is advisable to enter the areas on a low frequency range. In most of the test searches in this study, 5—8 no-peak areas were entered for one search, although the number depends on the spectral patterns of the unknown samples. In some earlier search systems, it is essential that no-peak areas do not overlap with the frequency ranges of input peaks; this is not a prerequisite for the proposed program.

The test search showed that the correct answers always obtain the highest

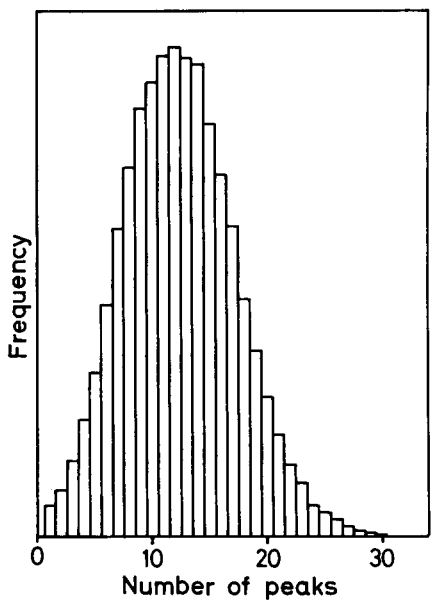


Fig. 4. Distribution of the number of peaks punched per spectrum in the ASTM cards between 1760 and 685 cm^{-1}. The ordinate scale is arbitrary.

scores, but that some noise spectra receive the same scores as the correct answers in several cases. The performance of the proposed search system was evaluated by the equations given by Erley [28]; the results obtained were an accuracy of 1.00, a precision of 0.88, and a figure of merit of 0.88. These values are higher than those of the FIRST-1 system, thus the performance of the proposed system seems rather better. However, it is difficult to compare the performance of various search systems rigorously, because the overall performance depends on numerous factors.

In some earlier systems, peak intensity data are taken into account in the computation of the scores, but they do not seem to be utilized so effectively as in the present system. Detailed descriptions of the earlier search programs have never been published, so that it is difficult to assess how the intensity data were utilized in these matching procedures. It appears that intensity data have been used to give larger weighting to stronger peaks in the computation of the scores, but the idea of penalty in eqn. (1) does not seem to have been utilized previously.

In the proposed method, input peaks are grouped into five intensity classes. It is possible to classify into ten or more groups, or to use observed transmittance or absorbance data in the scoring. However, when the difficulties involved in precise measurement of peak intensities and the dependence of relative peak intensities on experimental conditions are considered, such modifications would not give a better search performance than the present method.

The disadvantage of the present system seems to be its inapplicability to mixtures. Any search method based on intensity data will be ineffective in the identification of components in mixtures. For such purposes, it is necessary to rely on search methods based on different matching algorithms [30].

REFERENCES

1 B. R. Kowalski, P. C. Jurs and T. L. Isenhour, Anal. Chem., 41 (1969) 1945.
2 R. W. Liddel III and P. C. Jurs, Appl. Spectrosc., 27 (1973) 371.
3 R. W. Liddel III and P. C. Jurs, Anal. Chem., 46 (1974) 2126.
4 D. R. Preuss and P. C. Jurs, Anal. Chem., 46 (1974) 520.
5 H. B. Woodruff, S. R. Lowry and T. L. Isenhour, Anal. Chem., 46 (1974) 2150.
6 G. L. Grotch, Anal. Chem., 47 (1975) 1285.
7 G. L. Ritter, S. R. Lowry, H. B. Woodruff and T. L. Isenhour, Anal. Chem., 48 (1976) 1027.
8 H. B. Woodruff and M. E. Munk, Anal. Chim. Acta, 95 (1977) 13.
9 J. M. Comerford, P. G. Anderson, W. H. Snyder and W. H. Kimmel, Spectrochim. Acta, 33A (1977) 651.
10 J. C. Reid and E. C. Wong, Appl. Spectrosc., 20 (1966) 320.
11 R. O. Crisler, J. Am. Oil Chem. Soc., 45 (1968) 818.
12 K. Tanabe and S. Saëki, Anal. Chem., 47 (1975) 118.
13 L. A. Powell and G. M. Hieftje, Anal. Chim. Acta, 100 (1978) 313.
14 L. E. Kuentzel, Anal. Chem., 23 (1951) 1413.
15 ASTM, Codes and Instructions for Wyandotte-ASTM Punched Cards, Philadelphia, Pa., 1964.
16 T. Tamura, K. Tanabe, J. Hiraishi and S. Saëki, Jpn. Analyst, (1979) in press.

17 E. C. Penski, D. A. Padowski and J. B. Bouck, Anal. Chem., 46 (1974) 955.
18 A. Kondo and H. Nakagawa, Nihon Sekiyu Gakkaishi, 18 (1975) 291.
19 R. C. Fox, Anal. Chem., 48 (1976) 717.
20 R. A. Sparks, Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer, ASTM, 1964.
21 L. D. Smithson, Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using a 7090 Computer, Technical Documentary Report No. RTD-TDR-63-4265, Research and Technology Division, Wright-Patterson Air Force Base, 1964.
22 T. A. Entzinger, Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using a Honeywell-400 Computer, U.S. Public Health Service, Robert Taft Sanitary Engineering Center, 1964.
23 D. S. Erley, Fast Searching System for the Wyandotte-ASTM Infrared Data File, Chemical Physics Research Laboratory, The Dow Chemical Company, 1967.
24 D. H. Anderson and G. L. Covert, Anal. Chem., 39 (1967) 1289.
25 D. S. Erley, Anal. Chem., 40 (1968) 894.
26 L. H. Cross, J. Haw and D. J. Shields, Molecular Spectroscopy, The Institute of Petroleum, Adlard, Dorking (U.K.), 1968, 189 pp.
27 Yu. P. Drobyshev, R. S. Nigmatullin, V. I. Lobanov, I. K. Korobeinicheva, V. S. Bochkarev and V. A. Koptyug, Vestn. Akad. Nauk SSSR, 40 (1970) 75.
28 D. S. Erley, Appl. Spectrosc., 25 (1971) 200.
29 G. A. Massios, Amer. Lab., 3 (1971) 55.
30 R. W. Sebesta and G. G. Johnson Jr., Anal. Chem., 44 (1972) 260.
31 IRIS (Infrared Information System), University Computing Company, Palo Alto, Ca.
32 IRGO (Infrared Spectral Search Service), Singer Technical Services, New York, N.Y.
33 ISIS (Infrared Spectral Information System), Document No. LSR-98, Triangle Universities Computation Center, Research Triangle Park, N.C.
34 SIRCH (Infrared Spectral Data Retrieval System), ASTM Headquarters, Philadelphia, Pa.
35 S. Shimizu, IR-SPAN, Proc. Annual Meeting Chem. Soc. Jpn., 1 (1973) 2F14.
36 S. Kihara, Proc. Tokyo Conf. Appl. Spectrom., 9 (1973) 1B11.
37 J. Zupan, D. Hadzi and M. Penca, Comp. Chem., 1 (1976) 71.
38 R. N. Jones, Chem. Canada, 29 (1977) 23.
39 J. A. deHaseth, H. B. Woodruff and T. L. Isenhour, Appl. Spectrosc., 31 (1977) 18.
40 D. G. Cameron, SPIR, Operational Instructions, NRCC, Ottawa.

Analytica Chimica Acta, 112 (1979) 219—231
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

A COMBINED FORWARD—REVERSE LIBRARY SEARCH SYSTEM FOR THE IDENTIFICATION OF LOW-RESOLUTION MASS SPECTRA

J. KWIATKOWSKI and W. RIEPE*

*Institut für Spektrochemie und angewandte Spektroskopie***, Postfach 778, D-4600 Dortmund 1 (G.D.R.)*

(Received 23rd February 1979)

SUMMARY

A combined forward—reverse library search routine for low-resolution mass spectra is described. The routine requires binary-coded spectra. Masses and peak intensities are used for spectral comparison. On the basis of three possible search strategies, this routine is adaptable to analytical problems. The program was tested for 25 000 spectra from the ISAS, MSDC and EPA mass spectra libraries. The program is written completely in FORTRAN IV.

For the interpretation and identification [1—3] of mass spectra with the aid of a computer, three general methods are used: library search [4], pattern recognition [5] and artificial intelligence [6]. Library search has become the most widely applied method for the identification of mass spectra; the wide application of this technique is due to its general applicability, low cost and the availability of large spectral libraries. Furthermore, integrated data systems allow spectroscopists to create libraries for their own particular purposes.

Library searches can be done in the forward [6, 7] or in the reverse [8] direction. Attempts have been made to maximize the reliability of identification of an unknown which is definitely known to be in the library; in such cases, the reverse search method has been applied with some success [8]. In this case, the question asked is: "is this unknown a particular compound X?"; therefore, the library spectrum of compound X is used as the basis of comparison with the unknown spectra. In forward searches, the unknown is the reference for spectral comparison. The major advantage of the reverse search approach over the conventional forward search is that reliable identification of a compound may be possible even when the spectrum is rather severely distorted, e.g. by background or unresolved components. Because of this advantage, reverse search should be used to identify the spectra of a g.c.—m.s. run by library searches. In such cases, and in all other cases where a pure unknown spectrum cannot be expected, a reverse search method should be used.

**Supported by the Ministry of Science of the Land Northrhine-Westphalia.

This paper deals with a library search method which can be implemented on small computers. The computer used for spectrometer control can also be used for identification of spectra. Three different search strategies are possible with the same search algorithm: forward search, reverse search and combined forward–reverse search. Each strategy has its own advantages, so that this library search method is adaptable for many analytical problems.

Mass spectra libraries

Three libraries of experimentally measured low-resolution mass spectra were used for this study. The spectra of these libraries are base-peak normalized. Match results would be the same if sum-normalized spectra were used. The sources of spectral data were the MSDC mass spectra library, the EPA mass spectra libraries and the ISAS mass spectra library (see Table 1). The molecular weight distributions are given in Fig. 1. It should be pointed out that the search algorithm was applied to the libraries without any correction, so as to realize "practical" spectra including effects of impurities and other disturbances.

DESCRIPTION OF THE SEARCH ALGORITHM

In library search methods, an unknown spectrum is compared sequentially with the spectra in a library; the result of a single comparison is expressed as a match factor indicating the similarity of two spectra. A complete library search results in a list of spectra correlated to the unknown spectrum (hit list). The user may expect to find spectra useful for interpreting the unknown spectrum among the top n members in the hit list.

The search algorithm proposed here uses one-bit abbreviated spectra [9]. In one-bit coding, a spectrum is expressed as a binary string of zeroes and

TABLE 1

Information about mass spectral libraries

	ISAS[a]	MSDC[b]	EPA I[c]	EPA II[c]
Number of compounds containing only				
C, H	11	1968	397	1011
C, H, O	179	2257	2911	3012
C, H, N	55	313	409	441
Total number of spectra	600	6800	9960	9650

[a]This collection contains mass spectra recorded by a CH7 mass spectrometer at the Institut für Spektrochemie und angewandte Spektroskopie, Bunsen-Kirchhoff-Straße 11, D-4600 Dortmund 1. The library is currently being expanded and a paper is in preparation.
[b]Mass Spectrometry Data Centre, AWRE, Aldermaston, Berkshire, United Kingdom.
[c]From Varian MAT, Bremen.

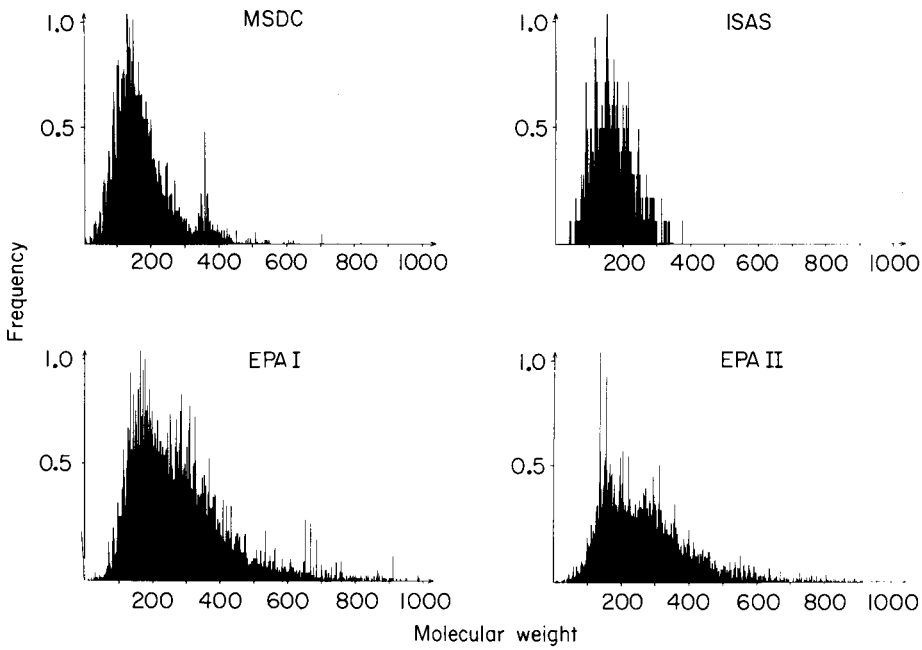


Fig. 1. Molecular weight distribution of EPA I, EPA II, MSDC and ISAS mass spectra libraries.

ones, zeroes indicating the absence of a mass peak and ones indicating the presence of a peak at a given mass number. The mass value is correlated to the bit position in the binary string [10]. The result of a spectrum comparison is also a binary string, which depends on the search technique selected. The number of ones in this string indicates the difference between the unknown and library spectra. Thus, with the reverse search technique, for example, all those bit positions which exhibit a one in the library spectrum and a zero in the unknown spectrum, are set to one. Thus a difference is indicated in the result string if a peak is present in the library spectrum but absent from the unknown spectrum. Analogously, no difference between the unknown and library spectra is indicated if the library spectrum is completely included in the unknown spectrum. Additional peaks (impurities and other artefacts) of the unknown spectrum will not be noted as "failures" as is required for reverse library searches. The resulting binary strings for reverse, forward and combined forward–reverse search techniques are shown in Table 2.

The unnormalized dissimilarity index is defined as the number of ones in the resulting binary string. This number can be used as a direct measure of the dissimilarity between the library and the unknown spectra. The best fit is obtained when the number of agreements between the two spectra is maximal. To obtain comparable dissimilarity indices for comparison of spectra, however, the number of ones in the resulting binary string must be normalized.

TABLE 2

Result of matching procedure

	Binary string			
Library spectrum	0	1	0	1
Unknown spectrum	0	1	1	0
Result:				
Reverse search	0	0	0	1
Forward search	0	0	1	0
Combined forward—reverse search	0	0	1	1

The normalized dissimilarity index (DI) is calculated from the formula

$$DI = N_{DA}/N_{DP} \quad (1)$$

where N_{DA} is the actual number of disagreeing bit positions in the binary spectra strings, and N_{DP} is the number of possible disagreeing bit positions in binary spectra strings.

The dissimilarity indices control the sequence of the entries in the hit list. The hit list provides hints for interpretation of the unknown spectrum which is aided by the values of the dissimilarity indices and the correlation of the proposed members of the hit list to the unknown spectrum. A hit list should answer the following questions. Is an identical spectrum present in the library? Are similar spectra present in the library to give hints for interpreting the unknown spectrum? Is no spectrum present in the library to give useful hints for interpreting the unknown spectrum?

The hit lists achieved by some library search systems are often inadequate for answering these questions. The result of the best fit is normalized to one, for example, and the dissimilarity indices of the following members of the hit list do not offer satisfactory answers to the above questions. Many systems do not give suitable answers to the third question. An answer to this question may seem trivial, but to have no proposals in the hit list is definitely preferable to having misleading proposals which may affect interpretation.

Figure 2 shows representative graphs of the hit list obtained by the search system developed. Parts 1a and 1b show the graphs obtained if there is an identical spectrum in the library. Identical spectra have very low dissimilarity indices, indicating a good fit. The unidentical spectra are separated decisively from the identical spectra. If there are no identical spectra in the library (Fig. 2, 2) the match factors for similar spectra are usually higher than for identical spectra but the slope of the graph is similar to the graph in parts 1a and 1b. If the library does not contain any spectrum giving useful hints for interpreting the unknown (Fig. 2, 3), the dissimilarity indices are consistently high, and there is no step in the graph separating the relevant and irrelevant parts of the hit list. This is important because it can be used in calculating a new weighted hit list from the hit lists shown in Fig. 3 (see below).

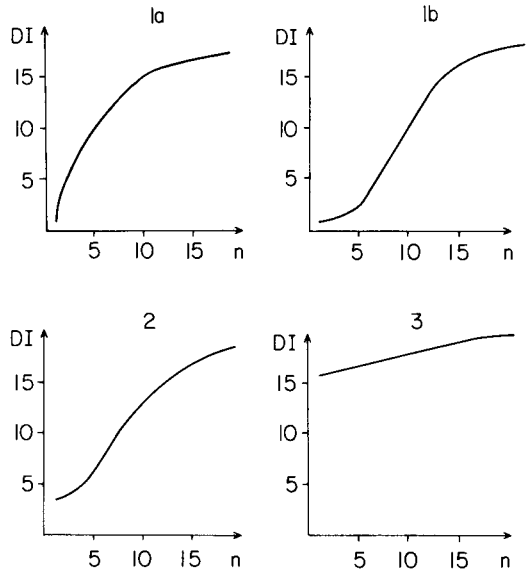


Fig. 2. Correlation between the top n members of the hit list and the dissimilarity index *DI*. (1a) One identical spectrum in the library; (1b) more than one identical spectrum in the library; (2) similar spectra in the library; (3) no similar spectra in the library.

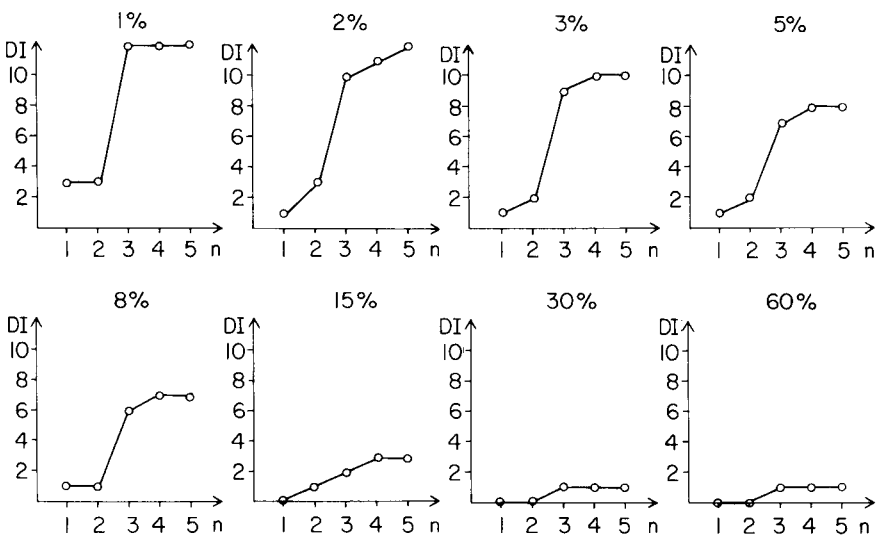


Fig. 3. Correlation between top n members of the hit list, the dissimilarity index (*DI*) and the intensity level.

The disadvantage of library searches based on one-bit coded mass spectra is that information about peak intensities is lost. This information is often necessary to distinguish between several spectra. From the point of view of

chemical analysis, peak intensities cannot be neglected because in most cases they contain important information concerning molecular structure. Even encoding mass spectra with two or more bits does not avoid this disadvantage entirely. Because of the importance of peak intensities, the hit list must be examined in relation to the intensity level. The graphs for several intensity levels are shown in Fig. 3; these plots belong to category 1a of Fig. 2. Obviously there is a range of intensity levels where the significance of the hit list is maximal, i.e. where the dissimilarity indices of relevant spectra are clearly separated by a step in the graph from irrelevant spectra. This ensures that the resulting hit list will not contain misleading proposals for interpreting the unknown spectrum.

To determine the range of intensity levels for reliable hit lists, the significance of the hit lists was examined for 100 representative samples at intensity levels of 0—70%. The significance of the hit lists was expressed by an empirically determined quality index which contains the following considerations: (1) is an unequivocal decision possible about an identical spectrum, similar spectra or irrelevant spectra in the library? (2) are the proposals of the hit list useful for interpretation of the unknown spectrum? (3) are there wrong proposals in the hit list?

At very low intensity levels the significance of the hit lists is poor because peaks with small intensities caused by noise or impurities falsify the dissimilarity index. High intensity levels will also lessen the significance, because only a few bits are set in the binary string and so there is no great power of discrimination for the library spectra. This is exemplified in Fig. 4, which shows the hit lists for the unknown (*p*-toluenesulfonamide) compared with the ISAS spectra library at several intensity levels, the spectra and the dissimilarity indices of the first proposals being given. It is never sufficient to calculate a single hit list at a particular intensity level. Statistical and random errors may cause unreliabilities. Such errors can be eliminated to a large extent if several single hit lists calculated for different intensity levels are combined. This combined hit list will nearly be free of the errors in the single lists, and so will be much more significant than the single lists. A theoretical basis for this suggestion involving information theory will be reported in a later paper.

The algorithm for merging single hit lists to form combined lists is shown in Table 3. The first step is to determine single hit lists in the intensity level range 1—8% (cf. Fig. 4). An interval of 1% is sufficient. Secondly, each single list is assigned to type 1, 2 or 3 as given in Fig. 2; the step in the graph for the hit list is also determined. The third step is a consistency test for all single hit lists: all single hit lists should belong to the same type (Fig. 2) and the step in the *DI—n* graph should occur at the same point. The result of this consistency test, i.e. type and height, are used to calculate a weighting factor for each single hit list. Fourthly, the weighted single hit lists are joined to form the combined hit list. To each proposal in the single hit list, a number is assigned which depends on its position in this particular list; the proposals up to the position where the step occurs are treated as if they all occupied the

first position. These numbers are multiplied by the weighting factor for each single list. To calculate the combined hit list, the numbers for the same proposals all over the single hit lists are added. This sum directly represents the match factor for the proposal in the combined hit list. This method of calculation ensures that if a proposal is always ranked highly in the single lists, it will be given a good match factor in the combined hit list. If, however, a proposal is ranked highly only once or twice, it will become less important in the combined list. This method of calculating the combined hit list is based on the principles shown in Figs. 2—4. The intensities of mass peaks are used indirectly in calculating the match factors for the combined hit list, and are selected in a range where they are important for identifying mass spectra. The intensity range and the numbers of intensity levels obviously depend on the library. In this work, these values were determined empirically. Calculation with the aid of information theory will be reported at a later date.

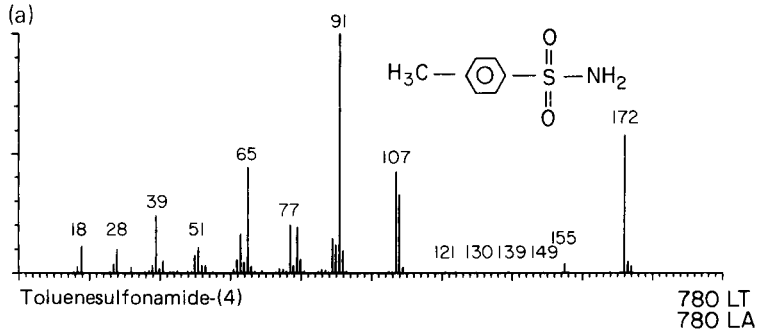
RESULTS AND DISCUSSION

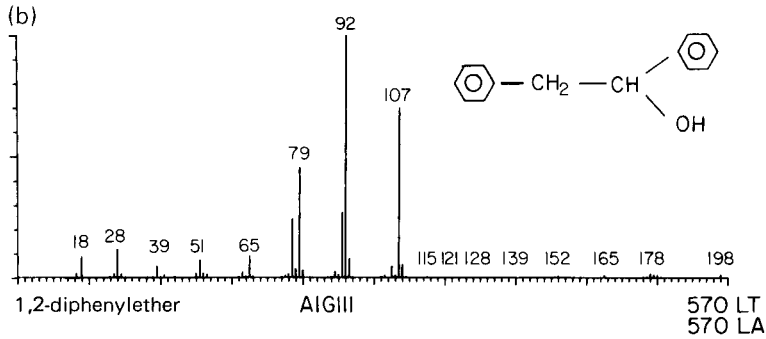
The proposed library search system was tested on spectra recorded by a CH7 Varian MAT mass spectrometer. The unknown spectra were matched against the ISAS, MSDC and EPA mass spectra libraries. Typical results are shown in Table 4. Table 4(a) shows results obtained by the forward search technique. As mentioned above, in this technique the basis of matching procedure is the unknown spectrum. Forward searches have some advantages if the unknown spectrum is not affected by impurities. In general, the library contains only fairly pure spectra, and in this case, the results produced by forward searching may be more valuable for prediction than those produced by reverse or combined forward—reverse searches.

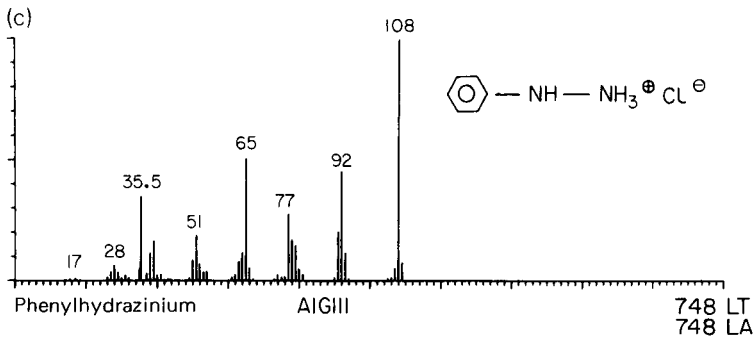
Results from the reverse search method are shown in Table 4(b). The advantage of this technique is that a spectrum can be recognized even when the unknown spectrum is only fairly pure. The reverse search technique should be used for g.c.—m.s. spectra or spectra resulting from mixtures by fractional evaporation [11, 12].

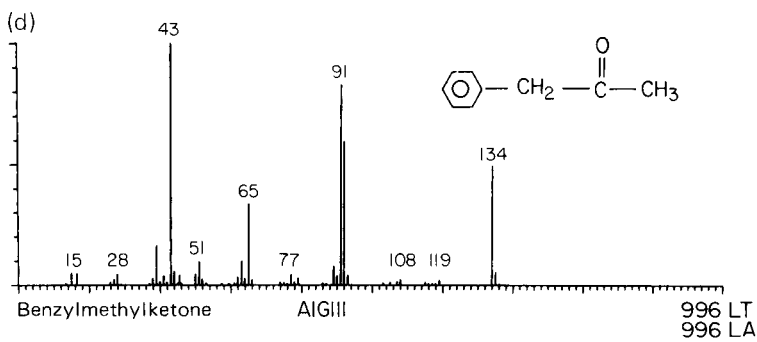
Table 5 shows the results of a reverse search applied to a mixture of acetamide and benzophenone. The composition of the mixture was varied and the different results were printed. With this method, impurities are neglected and the results of library searches are very similar whether the unknown spectrum is pure or impure. In the case of a unimolecular mixture of two components, both obtain the same match factor and are among the top members of the hit list.

Results produced by the combined forward—reverse search technique are shown in Table 4(c). This method gives the best possible combination of the advantages of reverse and forward searches. The results of the combined forward—reverse search are in general more useful than the results of separate reverse and forward searches. The main use of the combined technique is to search for an identical spectrum in the library. If such a spectrum is present,


(a)
91
H3C — S — NH2
O
O
172
65
107
39
77
18
28
51
121 130 139 149
155
Toluenesulfonamide-(4)
780 LT
780 LA


(b)
92
107
CH2 — CH
OH
79
18
28
39
51
65
115 121 128 139 152 165 178 198
1,2-diphenylether
AIGIII
570 LT
570 LA


(c)
108
NH — NH3⊕ CL⊖
65
92
35.5
77
51
17
28
Phenylhydrazinium
AIGIII
748 LT
748 LA


(d)
43
91
CH2 — C — CH3
O
134
65
15
28
51
77
108 119
Benzylmethylketone
AIGIII
996 LT
996 LA

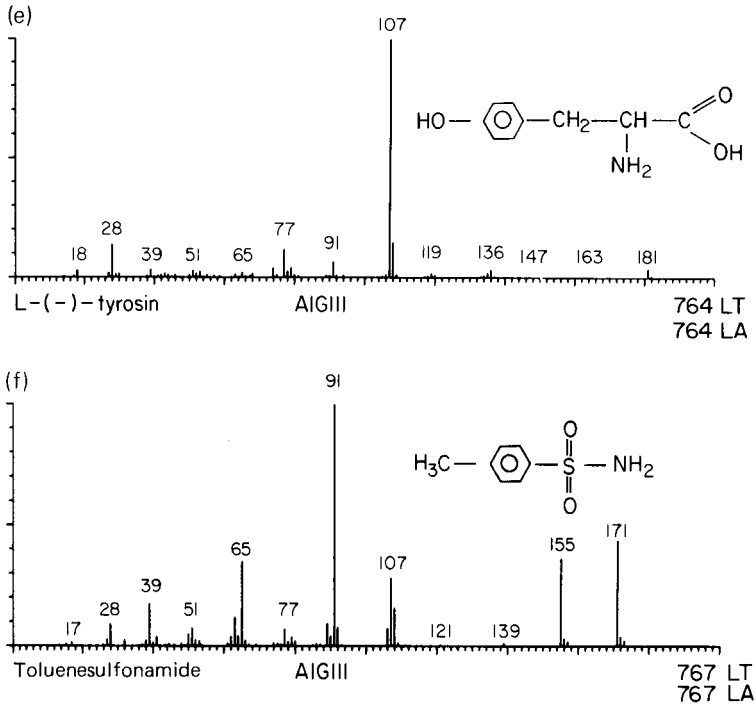


Fig. 4. Dissimilarity index as a function of intensity level for the unknown spectrum (f) and each of the reference spectra (a)—(e).

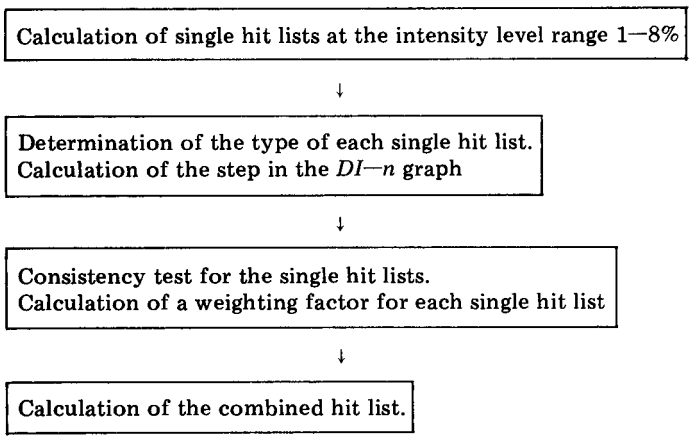
Reference spectrum	Intensity level (%)						
	0	1	5	10	30	40	70
	Dissimilarity index						
(a)	13	10	7	3	2	0	0
(b)	34	22	10	8	7	4	2
(c)	24	18	16	11	6	3	2
(d)	29	21	12	8	4	2	2
(e)	32	28	14	6	4	3	2

it is clearly and sharply indicated by the match factor, and a clear cut decision is possible about the presence or absence of an identical spectrum.

The proposed search system is capable of operating with forward, reverse and combined forward—reverse search techniques within the same program. Complex changes of algorithm, search strategy and handling of the system are unnecessary. The user can easily apply the three types of search strategy to the analytical problem at hand. The main disadvantage, that the peak intensities are ignored by search methods based on binary-coded spectra, is avoided by the new algorithm: several hit lists from search runs at different intensity levels are merged to form a combined hit list. Before combination,

TABLE 3

Flow chart for search routine



the single hit lists are examined for errors and checked by a consistency test, to ensure elimination of accidental errors. The number and values of the intensity levels were determined empirically here. A method of calculating the number and the values of the intensity levels by information theory will be reported later. These values are not fixed; they depend on the type and the number of the spectra in the library.

Program requirements

The program is written completely in FORTRAN IV. It was run on a MODCOMP IV computer. The complete program requires 24K core capacity. The time needed depends on the speed of the magtape and increases linearly with the size of the library.

The authors are indebted to Dr. D. Stüwer and Dr. R. M. Bombelka for their help in the preparation of the manuscript.

TABLE 4

Results of library searches
(I, Compound name; II, match factor; III, formula/molecular weight)

(a) FORWARD SEARCH

Hit list number	Unknown spectrum: benzophenone Reference library ISAS			Unknown spectrum: acetamide Reference library ISAS			Unknown spectrum: acetamide Reference library MSDC		
	I	II	III	I	II	III	I	II	III
1	Benzophenone	1.00	$C_{13}H_{10}O$ 182	Acetamide	1.00	C_2H_5NO 59	Acetamide	1.00	C_2H_5NO 59
2	Benzilic acid	0.98	$C_{14}H_{12}O_3$ 228	3-Ethoxypropionitrile	0.80	C_5H_9NO 99	3-Methoxypropylamine	1.00	$C_4H_{11}NO$ 89
3	Azobenzene	0.89	$C_{12}H_{10}N_2$ 182	Guanosine	0.75	$C_{10}H_{13}N_5O_5$ 283	Dichloroacetic acid	0.95	$C_2H_2O_2Cl_2$ 128
4	2-Benzoylbenzoic acid	0.65	$C_{14}H_{10}O_3$ 226	3-Heptanol	0.65	$C_7H_{16}O$ 116	Butyramide	0.75	C_4H_9NO 87
5	Benzhydrol	0.64	$C_{13}H_{12}O$ 184	1-Amino-2-propanol	0.64	C_3H_9NO 75	Glycol-dimercaptoacetate	0.75	$C_6H_{10}O_4S_2$ 210

(b) REVERSE SEARCH

Hit list number	Unknown spectrum: benzophenone Reference library ISAS			Unknown spectrum: acetamide Reference library ISAS			Unknown spectrum: benzophenone Reference library EPA I		
	I	II	III	I	II	III	I	II	III
1	Benzophenone	1.00	$C_{13}H_{10}O$	Acetamide	1.00	C_2H_5NO 59	Benzophenone	1.00	$C_{13}H_{10}O$ 182
2	Naphthalene	1.00	$C_{10}H_8$ 128	Urea	0.77	CH_4N_2O 60	Benzil	0.92	$C_{14}H_{10}O_2$ 210
3	Benzil	0.75	$C_{14}H_{10}O_2$ 210	Acetaldehyde	0.30	C_2H_4O 44	Azobenzene	0.75	$C_{12}H_{10}N_2$ 182
4	4-Benzoylpyridine	0.55	$C_{12}H_9NO$ 183	1,2-Diammonium-ethane dichloride	0.22	$C_2H_{10}Cl_2N_2$ 133	Azobenzene	0.75	$C_{12}H_{10}N_2$ 182
5	Azobenzene	0.40	$C_{12}H_{10}N_2$ 182	[a]			Benzidine	0.70	$C_{12}H_{12}N_2$ 184

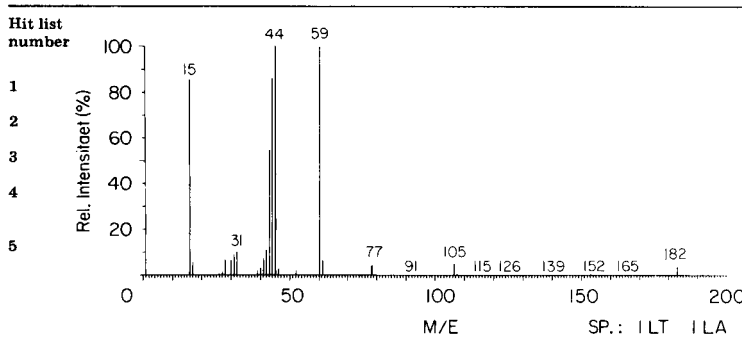
(c) COMBINED FORWARD—REVERSE SEARCH

Hit list number	Unknown spectrum: benzophenone Reference library ISAS			Unknown spectrum: acetamide Reference library ISAS			Unknown spectrum: benzophenone Reference library MSDC		
	I	II	III	I	II	III	I	II	III
1	Benzophenone	1.00	$C_{13}H_{10}O$ 182	Acetamide	1.00	C_2H_5NO 59	Benzophenone	1.00	$C_{13}H_{10}O$ 182
2	[a]			[a]			1-Phenyl-1,2-propanedione	0.21	$C_9H_8O_2$ 148
3–5	[a]			[a]			[a]		

[a]No other proposals with a match factor greater than 0.2.

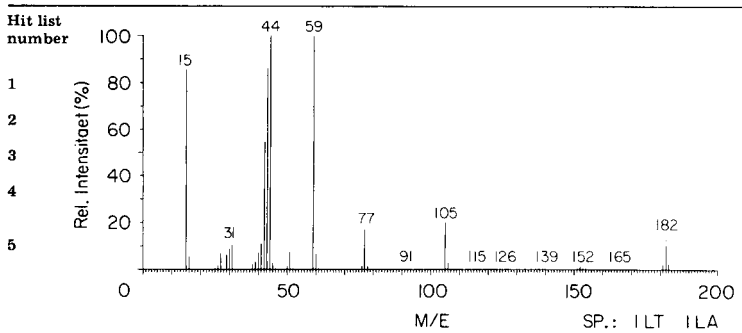
TABLE 5

Results for a compound mixture by the reverse search technique
(For I, II, III, see Table 4. Only the ISAS library was used.)



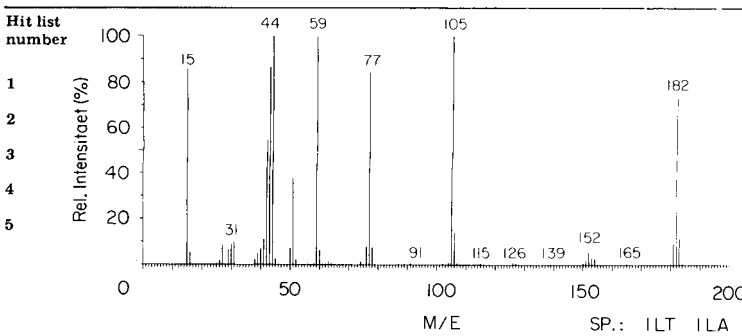
Acetamide with a small impurity of benzophenone

Unknown spectrum:		
I	II	III
Acetamide	1.00	C_2H_5NO 59
Urea	0.75	CH_4N_2O 60
Acetaldehyde	0.32	C_2H_4O 44
Ethylammonium chloride	0.25	C_2H_8ClN 81
a		



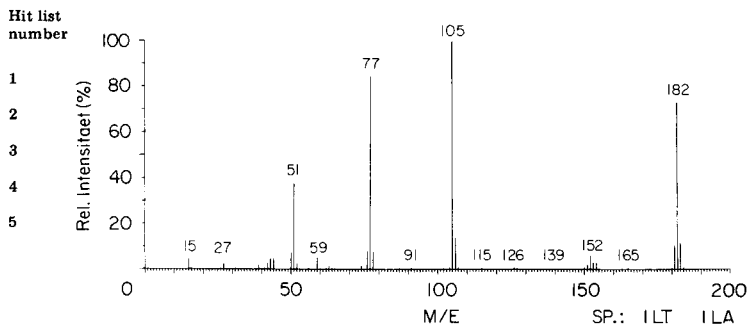
Acetamide with an impurity of benzophenone

Unknown spectrum:		
I	II	III
Acetamide	1.00	C_2H_5NO 59
Urea	0.75	CH_4N_2O 60
Acetaldehyde	0.31	C_2H_4O 44
Ethylammonium chloride	0.24	C_2H_8ClN 81
1,2-Diammonium-ethane dichloride	0.22	$C_2H_{10}Cl_2N_2$ 133



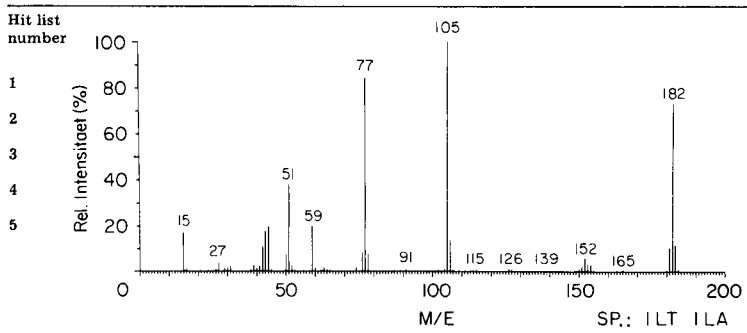
Mixture of acetamide and benzophenone (1:1)

Unknown spectrum:		
I	II	III
Acetamide	1.00	C_2H_5NO 59
Benzophenone	1.00	$C_{13}H_{10}O$ 182
Benzil	0.50	$C_{14}H_{10}O_2$ 210
Urea	0.47	CH_4N_2O 60
a		



Benzophenone with an impurity of acetamide

Unknown spectrum:		
I	II	III
Benzophenone	1.00	$C_{13}H_{10}O$ 182
Benzil	0.78	$C_{14}H_{10}O_2$ 210
4-Benzoylpyridine	0.36	$C_{12}H_9NO$ 183
Acetamide	0.31	C_2H_5NO 59
Benzilic acid	0.23	$C_{14}H_{12}O_3$ 228



Benzophenone with a small impurity of acetamide

Unknown spectrum:		
I	II	III
Benzophenone	1.00	$C_{13}H_{10}O$ 182
Benzil	0.67	$C_{14}H_{10}O_2$ 210
4-Benzoylpyridine	0.45	$C_{12}H_9NO$ 183
Azobenzene	0.32	$C_{12}H_{10}N_2$ 182

[a]No other proposals with a match factor greater than 0.20.

REFERENCES

1 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1464.
2 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 690.
3 J. Lederberg, in G. R. Waller (Ed.), Biochemical Application of Mass Spectrometry, Wiley, New York, 1972.
4 F. Erni and J. T. Clerc, Helv. Chim. Acta, 55 (1972) 489.
5 K. Varmuza, Fresenius Z. Anal. Chem., 268 (1974) 352.
6 L. E. Wangen, W. S. Woodward and T. L. Isenhour, Anal. Chem., 43 (1971) 1605.
7 K. Knock, R. Venkataraghavan and F. W. McLafferty, J. Am. Chem. Soc., 95 (1973) 4185.
8 F. P. Abramson, Anal. Chem., 47 (1975) 45.
9 S. L. Grotch, Anal. Chem., 43 (1971) 1362.
10 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
11 J. Franzen, H. Küper and W. Riepe, Anal. Chem., 46 (1974) 1683.
12 H. Hillig, W. Riepe and J. Kwiatkowski, Org. Mass Spectrom., 9 (1974) 1039.

Analytica Chimica Acta, 112 (1979) 233—243
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CALCULATION OF THE INFORMATION CONTENT OF RETRIEVAL PROCEDURES APPLIED TO MASS SPECTRAL DATA BASES

GEERT VAN MARLEN*

Department of Analytical Chemistry, Delft University of Technology, Jaffalaan 9, Delft (The Netherlands)

AUKE DIJKSTRA and HENK A. VAN'T KLOOSTER

Department of Analytical Chemistry, State University of Utrecht, Croesestraat 77A, Utrecht, (The Netherlands)

(Received 17th April 1979)

SUMMARY

A procedure has been developed for estimating the information content of retrieval systems with binary-coded mass spectra, as well as mass spectra coded by other methods, from the statistical properties of a reference file. For a reference file, binary-coded with a threshold of 1% of the intensity of the base peak, this results typically in an estimated information content of about 50 bits for 200 selected m/z values. It is shown that, because of errors occurring in the binary-coded spectra, the actual information content is only about 12 bits. This explains the poor performance observed for retrieval systems with binary-coded mass spectra.

In recent years, information theory has been applied in different fields of analytical chemistry. The information content has been introduced as an optimization criterion in thin-layer chromatography [1], gas chromatography [2, 3], infrared [4, 5] and mass spectrometry [6—9]. It has been shown that a mass spectrum even with binary-coded intensities still provides a large amount of information [6]. The information content is diminished considerably by correlations between spectral features [9]. Similar observations have been made for binary-coded infrared spectra [4, 5].

The influence of errors and matching criteria on the retrieval of binary-coded mass spectra has been discussed [9]. It was concluded that the performance of the retrieval primarily depends on the extent of errors occurring in the coded spectra and is hardly affected by the matching criterion used.

In this paper, a method is described for calculating the information content for binary-coded mass spectra as well as spectra coded by other means. In addition, a new algorithm for feature selection is presented. Finally, an approach is outlined for prediction of the performance of a forward search system with binary-coded spectra, when data bases which differ with respect to the number and the nature of the compounds involved are considered.

EXPERIMENTAL

The six mass spectral reference files used in this study are listed in Table 1. The spectra were binary-coded for a mass range of m/z 1—300, by using an intensity threshold of $T\%$ of the intensity of the base peak, with T varying from 0.1 to 20%. For some reference files, a 2-bit code was generated with three intensity levels, viz. 1, 5 and 20%, thus specifying peak intensities as: no peak, small peak, medium size and large peak, respectively.

For 1650 chemical compounds, duplicates were extracted from reference file F to enable an investigation of the effects of errors.

For the development and testing of the algorithms, a PDP11/45 minicomputer was used whereas the final computer programs were run on the IBM 370/158 computer system of the Delft University of Technology. All programs were written in FORTRAN IV.

INFORMATION CONTENT

Discontinuous distribution

Retrieval from a reference file containing N spectra each with n coded features (in this case m/z values) yields an amount of information called the "information content" of the retrieval procedure [9]. The information content for feature j, $I_H(j)$ in bits, without taking experimental and coding errors into account, is given by Shannon's equation [10]

$$I_H(j) = -\sum_{i=1}^{m} p_j(i) \log_2 p_j(i) \qquad (1)$$

where m represents the number of discrete values i for feature j with corresponding probabilities $p_j(i)$.

For binary-coded intensities, only two values for feature j are distinguished: either below or above a given threshold, coded as "0" or "1", respectively.

TABLE 1

Mass spectral reference files used

Ref.	Code-name	Number of spectra	Origin
A	ms9628	9628	MSDC file, release 1971[a]
B	msalkane	195	Alkane Spectra from A
C	ms15796	15796	MSDC file, release 1973[a]
D	ms02408	2408	Hydrocarbon Spectra from C
E	ms11346	11346	EPA/NIH file, release 1975[b]
F	ms22349	22349	Mix of C and E

[a]Mass Spectrometry Data Centre, Aldermaston, Gt. Britain.
[b]EPA/NIH Mass Spectral Data Base, 1975 edn., National Technical Information Service, Dept. of Commerce, 5285 Port Road, Springfield, Virginia 22151, U.S.A.

Since $\Sigma_{i=1}^{m} p_j(i) = 1$, it is obvious that in this case $p_j(0) = 1 - p_j(1)$. Equation (1) then reduces to

$$I_H(j) = -p_j \log_2 p_j - (1 - p_j) \log_2 (1 - p_j) \quad (2)$$

with $p_j = p_j(1)$. As an illustration for a large reference file, the probabilities and information contents for a number of m/z values, calculated with eqn. (2), are compiled in Table 2.

If the spectral features are considered to be independent, the total information content can be calculated with

$$I_H(1, 2 \ldots n) = \sum_{j=1}^{n} I_H(j) = -\sum_{j=1}^{n} \sum_{i=1}^{m} p_j(i) \log_2 p_j(i) \quad (3)$$

If there is a dependence, eqn. (3) has to be replaced by

$$I_H(1, 2 \ldots n) = -\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \ldots \sum_{i_n=1}^{m} p(i_1, i_2 \ldots i_n) \log_2 p(i_1, i_2 \ldots i_n) \quad (4)$$

The total number of probabilities to be estimated amounts to m^n. For a small number of features n and a relatively large number of spectra, eqn. (4) can be used to calculate the information content. In order to predict the information content of retrieval procedures for very large files from the statistical properties of small files, reliable estimates of p must be available. However, in practice the number of spectral features is large (for mass spectra a few hundred m/z values) and therefore it will be impossible to obtain an adequate estimate of

TABLE 2

Influence of the probability of occurrence p and the mismatch probability p_d on the information content I_H(Shannon, eqn. 2) and I'_H(Shannon, eqn. 14), respectively, for some binary-coded m/z values from reference file F (in bits)

m/z	Threshold level 1%				Threshold level 2%			
	p	I_H	p_d	I'_H	p	I_H	p_d	I'_H
27	0.44	0.99	0.24	0.20	0.40	0.97	0.22	0.22
28	0.41	0.98	0.26	0.16	0.37	0.95	0.25	0.16
29	0.42	0.98	0.19	0.28	0.38	0.96	0.18	0.28
42	0.52	1.0	0.12	0.47	0.45	0.99	0.12	0.46
44	0.47	1.0	0.17	0.34	0.39	0.96	0.16	0.34
51	0.52	1.0	0.14	0.41	0.44	0.99	0.10	0.50
53	0.52	1.0	0.15	0.39	0.42	0.98	0.12	0.45
55	0.58	0.98	0.11	0.48	0.51	1.0	0.10	0.53
56	0.42	0.98	0.10	0.12	0.36	0.94	0.10	0.49
57	0.51	1.0	0.12	0.47	0.43	0.98	0.10	0.52
65	0.43	0.98	0.11	0.49	0.35	0.93	0.09	0.51
69	0.50	1.0	0.09	0.56	0.44	0.99	0.08	0.58
77	0.56	0.99	0.09	0.55	0.49	1.0	0.09	0.17
75	0.43	0.98	0.10	0.52	0.35	0.93	0.08	0.53
91	0.44	0.99	0.10	0.52	0.38	0.96	0.07	0.58

the p values even with the sizes of the spectral data bases which are presently available. Application of eqn. (4) then leads to a maximum information content of $\log_2 N$ bits.

Continuous distributions and correlations

When the discontinuous distributions of the probabilities p_j in eqn. (1) can be approximated by a continuous normal distribution, the sum in eqn. (1) is replaced by the integral

$$I_G(j) = -\int_{-\infty}^{+\infty} p_j(x) \log_2 p_j(x)\, dx \tag{5}$$

with $p_j(x)$ as the Gaussian distribution function for feature j with value x measured in histogram units and $I_G(j)$ the information content for the integral form. After integration eqn. (5) becomes

$$I_G(j) = \tfrac{1}{2} \log_2 2\pi e\sigma_j^2 \tag{6}$$

where σ_j^2 is the variance of the normally distributed feature j. Thus the information content is a logarithmic function of this variance. The n-dimensional equivalent becomes

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2(2\pi e)^n |\mathrm{COV}| \tag{7}$$

when I_G is the total information content for n features and $|\mathrm{COV}|$ the determinant of the covariance matrix, filled with the variances σ_i^2 and the covariances σ_{ij} [9]. Combination of eqns. (6) and (7) finally results in

$$I_G(1, 2 \ldots n) = \sum_{j=1}^{n} I_G(j) + \tfrac{1}{2} \log_2 |\mathrm{COR}| \tag{8}$$

with $|\mathrm{COR}|$ as the correlation determinant defined as

$$\mathrm{COR} = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \ldots\ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots\ldots\ldots\ldots & & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \ldots\ldots & \rho_{3n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \rho_{n1} & \ldots\ldots & \ldots\ldots & \ldots\ldots & 1 \end{vmatrix}$$

The correlation coefficient ρ_{ij} between the features i and j can be calculated with the estimates of the variances σ_i^2 and σ_j^2 and the covariance σ_{ij} [9] from the equation $\rho_{ij} = \sigma_{ij}/(\sigma_i^2\, \sigma_j^2)^{\frac{1}{2}}$.

The second term in eqn. (8) is a correction of the information content caused by the interdependence of the features involved.

Binary-coded spectra and feature selection

Although binary-coded features can hardly be considered normally distributed, one can make an estimate of variances, covariances and correlation coefficients. The following equations can be derived:

$$\sigma_j^2 = N\,p_j\,(1-p_j)/(N-1) \tag{9}$$

$$\sigma_{ij} = (p_{ij} - p_i p_j)/[p_i p_j (1-p_i)(1-p_j)]^{\frac{1}{2}} \tag{10}$$

with p_i and p_j being the probabilities of coding feature i or j, respectively, present ("1") and p_{ij} the probability of coding both features present. Application of eqns. (2), (6) and (9) yields the information contents I_H and I_G as a function of p_j for one binary feature j. The results are presented in Fig. 1. The approximation of I_H by I_G for values of p between 0.15 and 0.85 results in a maximum error of 0.05 bit. Instead of making a correction in eqn. (8) for the differences between I_H and I_G [4, 9], the total information content is calculated from

$$I_G(1, 2 \ldots n) = \sum_{j=1}^{n} I_H(j) + \tfrac{1}{2} \log_2 |\mathrm{COR}| \tag{11}$$

The effect of the correction for correlated features (the second term in eqn. 11 is best illustrated in Table 3. In this Table the information contents calculated with eqns. (3), (4) and (11) are given for a set of m/z values with low correlation derived from reference file A [9]. From these numbers, it appears that eqn. (11) gives a reasonable estimate of the total information content.

However, as shown in Table 4 for an artificial set of two highly correlated features, calculation of the information content with eqn. (11) will sometimes result in an overcorrection for the correlations between the features. In addition, the cumulative effect of the correlations will make it impossible to calculate the total information content for all the features concerned. To avoid this problem, features that tend to show the effect of overcorrection must be deleted during the calculation. This is achieved by applying a feature selection

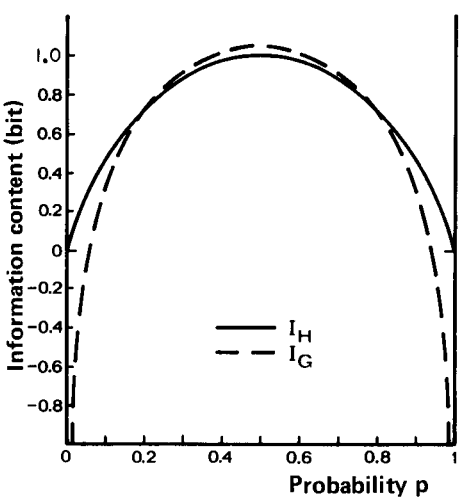


Fig. 1. The information contents I_H (eqn. 2) and I_G (eqns. 6, 9) as a function of the probability p for a single binary-coded feature.

TABLE 3

Information contents for a set of n poorly correlated m/z values derived from reference file A (in bits)

n	m/z value	$\Sigma I_H(j)$[a]	$I_H(1, 2 \ldots n)$[b]	$I_G(1, 2 \ldots n)$[c]
1	77	1.0	1.0	1.0
2	69	2.0	2.0	2.0
3	27	3.0	3.0	3.0
4	50	4.0	3.8	3.8
5	45	4.9	4.6	4.7
6	40	5.9	5.4	5.5
7	57	6.9	6.2	6.2
8	75	7.8	6.8	7.0
9	81	8.7	7.4	7.7
10	44	9.7	8.1	8.4

[a]Sum of individual information contents, eqn. (3).
[b]Shannon information content, eqn. (4).
[c]Information content corrected for correlation, eqn. (11).

TABLE 4

Influence of a high correlation coefficient ρ on the information content for a few sets of two artificial features 1 and 2 (in bits)

$I_H(1)$[a]	$I_H(2)$[a]	ρ_{12}	$\Sigma I_H(j)$[b]	$I_G(1, 2)$[c]
1.0	1.0	1.0	2.0	$-\infty$[d]
1.0	1.0	0.9	2.0	0.8[e]
1.0	1.0	0.8	2.0	1.3
0.9	0.8	1.0	1.7	$-\infty$[d]
0.9	0.8	0.9	1.7	0.5[e]
0.9	0.8	0.8	1.7	1.0

[a]Shannon information content, eqn. (2).
[b]Sum of individual information contents, eqn. (3).
[c]Information content corrected for the correlation between 1 and 2, eqn. (11).
[d]Information content not defined (zero logarithm).
[e]Information content overcorrected for correlations.

algorithm, resulting in an information content $I_G(1, 2 \ldots l)$ for l selected features with l less or equal to n. A detailed description of the algorithm is given in the Appendix. These calculations have been carried out as a function of the intensity threshold T for a number of binary-coded mass spectral reference files. The information content I_G and the number of features selected are presented in Table 5.

Extension to non-binary-coded spectra

Since the method of calculating the information content uses only linear

TABLE 5

Information content $I_G(1, 2 \ldots l)$ and the number of selected m/z values l for the binary-coded reference files at 6 different threshold levels from eqns. (10) and (11) (in bits)

Threshold level	0.1%		1%		2%		5%		10%		20%	
Reference file	I_G	l	I_G	l	I_G	l	I_G	l	I_G	l	I_G	l
msalkane (B)	—		11	34	—		—		—		—	
ms9628 (A)	—		42	140	—		—		—		—	
ms02408 (D)	29	95	22	80	19	77	16	77	12	76	—	
ms15796 (C)	47	157	46	156	41	159	32	159	24	164	17	171
ms11346 (E)	55	194	55	196	51	198	40	197	29	188	20	197
ms22349 (F)	—		52	174	47	176	—		—		—	

correlations between the features, another estimate of these correlations may be obtained by applying the 2-dimensional expressions of eqns. (4) and (11) for the features 1 and 2:

$$I_H(1, 2) = -\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} p(i_1, i_2) \log_2 p(i_1, i_2) \tag{12}$$

$$I_G(1, 2) = I_H(1) + I_H(2) + \tfrac{1}{2} \log_2 (1 - \rho_{12}^2) \tag{13}$$

$I_H(1, 2)$ can easily be calculated with a large set of spectra, since only four p values have to be estimated for binary-coded features. Equating $I_H(1, 2)$ and $I_G(1, 2)$ then leads to a new estimate for the correlation coefficient ρ_{12}; this may then be used in eqn. (11), thus circumventing overcorrection. The same feature selection procedure mentioned above has been applied to obtain the information content I_G for the binary-coded reference files. Table 6 gives a brief example of the results.

This alternative manner of quantifying correlations between the features can also be used to assess the information content for non-binary-coded features, provided that the number of spectra is large enough to estimate all p values used in eqn. (12). For some reference files, the features have been converted to a 2-bit code; the information contents and the numbers of features selected are given in Table 7.

TABLE 6

Information content I_G (eqns. 11 and 13) and the number of selected m/z values, l, for the binary-coded reference files (threshold level 1%) (in bits)

Reference file	$I_G(1, 2 \ldots l)$	l	Reference file	$I_G(1, 2 \ldots l)$	l
msalkane (B)	13	48	ms15796 (C)	60	257
ms02408 (D)	23	93	ms11346 (E)	78	282
ms9628 (A)	49	214	ms22349 (F)	68	276

TABLE 7

Information content I_G and the number of selected m/z values l for 3 reference files, converted into a 2-bit code (eqns. 11, 12, 13) (in bits)

Reference file	$I_G(1, 2 \ldots l)$	l
ms02408 (D)	40	103
ms15796 (C)	110	288
ms11346 (E)	137	284

Influence of errors

For actual retrieval of spectra, the information content must be corrected for the uncertainty remaining after coding the spectra. The uncertainty in the coded spectra is caused by deviations in experimental, recording and coding conditions. Hence, for binary coded features, eqn. (2) must be replaced by:

$$I'_H = -p \log_2 p - (1-p) \log_2(1-p) + p[p_{1/0} \log_2 p_{1/0} + (1-p_{1/0}) \log_2(1-p_{1/0}$$
$$+ (1-p)\,[p_{0/1} \log_2 p_{0/1} + (1-p_{0/1}) \log_2(1-p_{0/1})] \quad (14)$$

with $p_{1/0}$ and $p_{0/1}$ being the probabilities of coding a "0" or a "1", when the spectrum of the reference compound contains a "1" or a "0", respectively, for that feature. From the probability of finding a mismatch between two spectra of the same compound (the "mismatch probability" p_d) the "error probabilities" $p_{1/0}$ and $p_{0/1}$ can be calculated from [11]

$$p_{1/0} = p_d/2p \text{ and } p_{0/1} = p_d/2(1-p)$$

The mismatch probabilities are estimated from a set of 1650 pairs of spectra of the same compound, extracted from the reference file F. The large influence of the mismatch and error probabilities on the information content actually obtained in a retrieval process is shown for some m/z values in Table 2. The mismatch probability ranges up to 0.26, with the maximum values in the low m/z area.

With the assumption that the error probabilities for the different features are independent, the correction of the information content can also be done by replacing $I_H(j)$ by $I'_H(j)$ in eqn. (11), yielding I'_G. In Table 8 the information contents I_G and I'_G, calculated for reference file F, binary-coded with two threshold levels, are given as an example.

Inspection of the results indicates that the information content drops dramatically from 52 to 11 bits and from 47 to 12 bits, respectively. This is much more than previously predicted [9]. Repeating the calculation for reference file A gave the same effect. This result explains the generally poor performance of a retrieval system based on binary-coded mass spectra [11].

CONCLUSIONS

The model described for the calculation of the information content of

TABLE 8

Information contents I_G (eqns. 10, 11) and I'_G (bits) (eqns. 10, 11, 14) as a function of the number of selected m/z values l for reference file F, binary-coded at two threshold levels

l	Threshold level 1%		Threshold level 2%	
	$I_G(1, 2 \ldots l)$	$I'_G(1, 2 \ldots l)$	$I_G(1, 2 \ldots l)$	$I'_G(1, 2 \ldots l)$
10	8.7	4.3	8.4	4.4
20	14.9	6.8	14.5	7.1
30	20.3	8.5	19.5	8.9
40	25.0	9.7	23.7	10.3
50	29.1	10.5	27.4	11.3
60	32.7	11.0	30.8	12.0
70	36.1	11.2	33.8	12.5
80	39.1	11.3	36.5	12.7
90	41.8	—	38.9	12.8
—				
—				
Max.[a]	51.7	11.3	47.4	12.8

[a]The maximum number of selected m/z values is 174, 77, 176 and 89, respectively.

retrieval procedures involves estimates of variances and correlation coefficients for a limited number of spectra in order to predict implicitly the probabilities for all possible codes. For features with extremely high correlations, the first model would lead to pessimistic estimates of the resulting information content. Use of a feature selection procedure is inevitable.

The second method for the calculation of the information content includes a different measure of the correlations (eqn. 13) which results in a contribution of almost all coded features to the information content. For this reason, the second method is expected to yield more reliable estimates of the information content. This method is also feasible for spectra coded by non-binary techniques, although it should be refined in order to deal with the influence of experimental and coding errors.

In the model discussed, the reference files are considered as select samples of complete populations of mass spectra. Consequently, in the design or evaluation of retrieval methods the reference file to be used for the calculation of the information content should be as representative of a certain population of compounds as possible. For example, the results for alkane spectra (reference file D), even without a correction of the information content for errors in the spectra, show clearly that binary coding does not yield enough information and is not recommended for retrieval purposes.

For retrieval from reference files containing spectra of a wider variety of compounds (such as C, E and F), binary coding is more promising. With a threshold of 1% of the intensity of the base peak, an information content of about 50 bits for 200 selected m/z values is typical. This value decreases rapidly above threshold values of 2%, mainly because of the diminishing number of peaks per spectrum.

However, most of the information is lost because of the errors that occur in the binary-coded spectra. The performance of a retrieval system based on the binary-coded reference file A confirms this conclusion [11]. The prediction that a threshold of 2% would yield a better retrieval performance than 1% is confirmed by the preliminary results obtained with reference file F.

The authors gratefully acknowledge valuable discussions with Louis van Norel, Peter Cley and Jan Van den Hende.

APPENDIX: FEATURE SELECTION ALGORITHM

To calculate the information content I_G for a near-optimal subset of l out of n features, eqn. (11) is converted to

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2 |S| \tag{a1}$$

with the elements of determinant S defined as $s_{ij} = 2I_H(i) \cdot 2I_H(j) \cdot \rho_{ij}$. This equation introduces a scaling which avoids computational problems such as arithmetical under- and over-flow. The diagonal elements of S range in value from 1 to 4 and represent the exponential form of the information content for every feature ($I_G(i) = \frac{1}{2} \log_2 s_{ii}$). The non-diagonal elements vary from -4 to $+4$, since the correlation coefficient ρ ranges from -1 to $+1$, and represent the "covariances" between the features. The information content is calculated after the covariance elements of S have been zeroed by using the Gaussian elimination method [12] with

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2 \prod_{i=1}^{n} s'_{ii} = \sum_{i=1}^{n} \tfrac{1}{2} \log_2 s'_{ii} \tag{a2}$$

where s'_{ii} is the "variance" element after correction for the correlations between the features.

Because of the effect of overcorrection for correlations between the features only l features contribute to the calculated information content I_G. These features are selected together with their contribution to I_G by the following procedure:

1. select the feature i with the highest value of s_{ii};
2. correct the variances of the remaining features k for the correlation between i and k with

$$s'_{kk} = s_{kk} - s_{ik}^2 / s_{ii} \tag{a3}$$

3. find the next feature n with the highest corrected value of s'_{nn};
4. correct the covariances of the remaining features k for the correlation between feature k and all features j previously selected, excluding feature n, with

$$s'_{nk} = s_{nk} - s_{jk} \cdot s_{jn} / s_{jj} \tag{a4}$$

(Note that for $j = n$, s'_{nk} is zero in which case s'_{nk} is not calculated.)

5. replace i by n and restart at step 2.

The iterative procedure is terminated when all features have been selected or when the highest corrected value of s'_{nn} in step 3 becomes less than 1; all remaining features k will then have s'_{kk} less than 1 and $\frac{1}{2} \log_2 s'_{kk}$ becomes negative.

The procedure described has been chosen to avoid the correction of all covariances in S; the elements need to be corrected only for the correlations with the features already selected. In programming this procedure the indices i, j, k and n are replaced by elements of a pointer array, in which the sequence of the features is stored, and which enables an indirect reference to the rows and columns of S without time-consuming rearrangements. The square matrix S is symmetric and only the upper triangle of the matrix need be stored, giving a considerable reduction in storage requirements. However, if possible the linear storage of S should be avoided as otherwise all indices in the eqns. (a3) and (a4) have to be computed with a special function. The calculation of the information content I_G for the mass spectral reference files in this study, with 300 binary-coded m/z values, requires the storage of 300 × 300 elements of the matrix S. The execution time depends linearly on the number of coded features multiplied by the number of features selected. On the IBM 370/158 computer system this amounts to approximately 40 s for 100 features selected and 300 m/z values coded.

REFERENCES

1 D. L. Massart, J. Chromatogr., 79 (1973) 157.
2 P. F. Dupuis and A. Dijkstra, Anal. Chem., 47 (1975) 379.
3 A. Eskes, P. F. Dupuis, A. Dijkstra, H. Declerc and D. L. Massart, Anal. Chem., 47 (1975) 2168.
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 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
7 F. Erni, thesis no. 4296, Eidgen. Techn. Hochschule, Zürich (1972).
8 F. Erni and J. T. Clerc, Helv. Chim. Acta, 55 (1972) 489.
9 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
10 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill., 1949.
11 G. van Marlen, A. Dijkstra and H. A. van't Klooster, Anal. Chem., 51 (1979) 420.
12 R. W. Hamming, Introduction to Applied Numerical Analysis, McGraw-Hill, New York, 1971, Ch. 5.

Analytica Chimica Acta, 112 (1979) 245—252
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

THE DARC PLURIDATA SYSTEM: THE ^{13}C-N.M.R. DATA BANK‡

J. E. DUBOIS* and J. C. BONNET
ITODYS, Université Paris VII, 1, Rue Guy de la Brosse, 75005 Paris (France)
(Received 3rd May 1978)

SUMMARY

The capabilities of the DARC system are discussed and illustrated by the storage and retrieval functions of the ^{13}C-n.m.r. data bank of the DARC PLURIDATA system. The data covered by the bank, as well as the input stream to the bank and validation of the spectra, are described. Particular stress is laid on the DARC structural retrieval system, which illustrates the interactive interrogation of a chemical data bank by means of the structural diagram of a molecule, i.e. the universal language in chemistry. The potential of the ^{13}C-n.m.r. data bank in computer-aided structural elucidation is outlined.

The DARC PLURIDATA system (DPDS) [1] currently implemented at the "Centre d'Information et de Documentation Automatique" (CIDA) of Paris VII University is an application of the DARC system for generalized chemical information to the field of chemical data banks; the DARC system [2, 3] is primarily concerned with structural chemical information. The ^{13}C-n.m.r. data bank is one of the many components of the DPDS. Out of the ten DPDS data banks currently grouped on a PDP-11/35 computer, three are fully operational and are accessible on-line through CYCLADES, the French computer network; these relate to ^{13}C-n.m.r. mass spectra, and x-ray diffraction data. All the DPDS banks are designed by using the same simple architecture: the information data base (IDB) containing numerical values, spectra and bibliographic data is associated with a structural data base (SDB) containing the storage and retrieval data relevant to the chemical structures of the data bank. The computer processing of the structure of chemical compounds is, of course, the heart of all chemical data banks, notably ^{13}C-n.m.r. data banks.

The ^{13}C-n.m.r. data bank as described here illustrates the capabilities of the DARC system for storage and retrieval of structural data. Particular stress is laid on demonstrating how the manager or user of the data bank can communicate with it in his natural language, i.e. the structural diagram of the chemical compound, regardless of the internal processing language. Internal processing languages, such as the topological description language of chemical compounds, support the communication language while remaining transparent to the user.

‡This paper was presented at the Conference on Computer Techniques and Optimization, Amsterdam, April, 1978.

Coverage by the data bank

The ^{13}C-n.m.r. data bank currently contains 8000 distinct compounds and 11000 spectra, along with their experimental conditions and the source of the data. Updating occurs at a rate of 500 structures and 600 spectra per month. Each chemical compound is stored once and only once in the data bank, but may be related to one or several spectra.

The data originate mainly from the literature, but also come from n.m.r. laboratories or from private collections. The literature indexing covers the following journals: Journal of the Chemical Society, Perkin Transactions 1 and 2, Journal of the American Chemical Society, Journal of Organic Magnetic Resonance, Journal of Organic Chemistry, Tetrahedron Letters, and Journal de Chimie Physique. Systematic scanning of this literature was started in January, 1976, and the back issues have been covered back to 1970. The input stream to the data bank offers the possibility of evaluating the actual contr bution of any other journal to the bank and, thereby, of assessing the interest that lies in adding it to the above-mentioned list.

DESCRIPTION OF THE INPUT STREAM TO THE DATA BANK

Indexing

Each article dealing with ^{13}C-n.m.r. is analyzed and, if it is selected, it is indexed to give two types of flow sheets: (a) a flow sheet for each chemical structure and its corresponding spectrum in the paper, which carries the structural diagram of the compound, its molecular formula, and the list of chemical shift values; and (b) a flow sheet which describes the source of the data (journal, author(s), title, name of the laboratory, etc.) and the experimental conditions cited therein. In flow sheet (a), the correspondence between the chemical shifts and the location of the carbon atoms within the structure is obtained by numbering these atoms in the list order of the chemical shifts.

Input of the flow sheets to the data bank (Fig. 1)

The originality of the input stream lies in the input route of the structural diagram to the SDB. This route is linked to the input route of the n.m.r. spectı to the IDB. Each chemical compound is recorded only once in the data bank. This function relies on the use of the DARC code, which is a unique and unambiguous topological description of the chemical structure. Since the DARC code is automatically generated, only those specific features which are required for the unambiguous assignment of the chemical shifts to the carbon atoms are discussed below.

Input of the chemical structure is carried out by using a specific device, the TOPOCODEUR, which was designed and built at the CIDA. The structural diagram is drawn from the TOPOCODEUR keyboard by moving a pointer on the screen from one point to another of the 16 × 10-point matrix. These points are thereby linked by the bonds of the formula; the atomic symbols of hetero-atoms, charges, abnormal valences, etc., are indicated at each appropriate point and the carbon atoms are labelled by their number on the

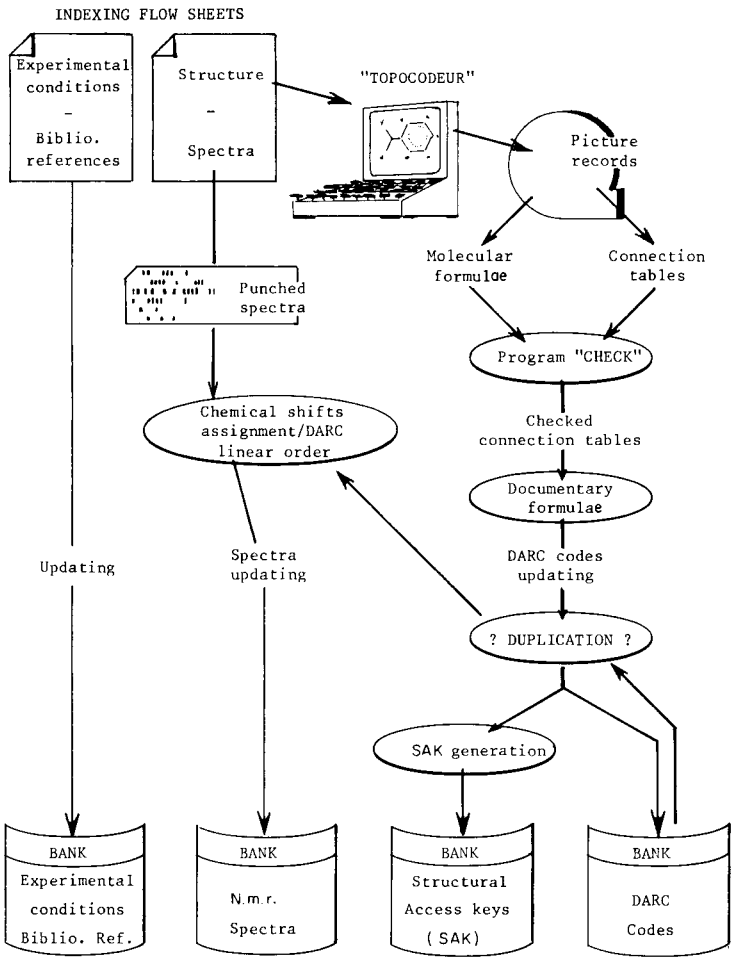


Fig. 1. General flow chart of the input stream to the data bank.

flow sheet. The status of the screen, i.e. the drawn picture, is then recorded on magnetic tape. This magnetic tape is processed by a set of computer programs which lead to the unique and unambiguous DARC code for each input structure.

If the answer to the question "is the code new?" is yes, then the input compound is recorded in the data base. If the answer is no, this means that the compound is already in the data base. Duplications are thus eliminated. However, it can be seen (Fig. 1) that the input stream includes a checking procedure which mainly consists in matching the input molecular formula reported on the flow sheet against the one computed from the drawn diagram. There is also a documentary formula procedure which consists in automatically solving the problem of multiple representation of a single compound and, in particular, the problems of

mesomerism and tautomerism; multiple but equivalent formulae given by author are automatically linked to a conventional formula, which is the principal input to the SDB for this compound. The SAK (structural access keys) generation procedure yields the set of computer data elements which are needed to solve efficiently the problem of structural retrieval. These structural access keys are stored and organized into the SDB along with the DARC code for every new compound

The input route of the spectra is as follows. The chemical shifts are punched onto cards in the order defined on the flow sheet, which is an arbitrary order related to the initial numbering of the carbon atoms. The unique and unambiguous DARC code results from an algorithm [4, 5] which assigns a unique label (Fig. 2), A_i or B_{ij}, to each atom, which is tantamount to a unique numbering ($FO \equiv 1$, $A_1 \equiv 2$, $A_2 \equiv 3$, $A_3 \equiv 4$, $B_{11} \equiv 5$, etc.). The structure is thus linearly ordered and this order is unique, i.e. a given structure will always receive the same final numbering. Therefore, by establishing the correspondence between the arbitrary input numbering of the carbon atoms and their unique final numbering, the assignment of chemical shifts to the carbons of a structure will be unique and unambiguous. Experimental conditions and bibliographic references enter the data bank directly from an interactive terminal.

This is an overall view of the main functions of the data input stream, i.e. the creation and updating of the data bank. Since, in every actual data bank,

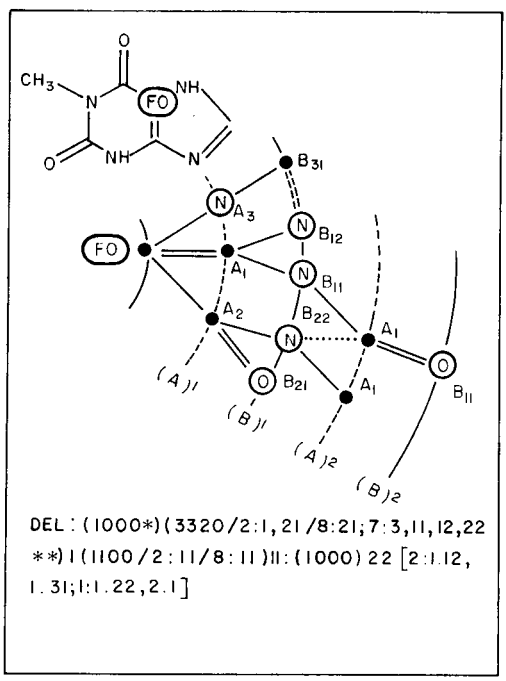


Fig. 2. Example of a structure (1-methylxanthine) ordered linearly by the DARC algorithn and its related unique and unambiguous DARC code (descriptor by environment limited to B DEL).

the data contained should be reliable and must be checked by validation procedures, the manner in which the spectra of the ^{13}C-n.m.r. data bank are validated merits discussion.

Validation of the spectra

In principle, every spectrum proposed for a given structure, either by a single source or by various sources, is a candidate for input into the data bank. Two types of validation procedure are used to check it: the first is integrated into the input stream and is systematic, whereas the second is periodical and concerns all the data stored in the bank at a given time.

The first procedure consists in matching the new spectrum proposed for an old compound against the spectra previously recorded for this compound. If this comparison shows a discrepancy in chemical shifts, the causes are sought manually. The discrepancy might be due to faulty indexing of the spectrum on the flow sheet or to faulty punching; in either case, the spectrum can be corrected. The discrepancy can also arise from a variation in experimental conditions in which case an n.m.r. expert must decide whether to keep or discard a spectrum.

The second procedure consists in checking the internal consistency of a dictionary which contains all the distinct chemical shifts in the bank and associates them with the environment of their corresponding carbon atom. This dictionary is progressively updated in parallel with the bank, and its periodical analysis is an efficient tool for spotting any abnormal values of chemical shifts.

RETRIEVAL

The retrieval function is the one that most concerns the user of the data bank. The ^{13}C-n.m.r. data bank can be interrogated by structure or substructure, by a complete or partial spectrum, by a bibliographic reference (author(s), year and journal), or by any combination of these criteria.

Here, particular stress is put on interrogation by chemical structure or substructure, because it illustrates the interactive interrogation of the chemical data bank by means of the chemist's natural language, i.e. by the structural diagram. This interrogation is one of the most important features of the DARC system [4]. Indeed, the question to the bank is formulated by simply drawing the structure or substructure onto the screen of the graphical terminal. If no such terminal is available, the drawing of the question is replaced by a simple description of the connection between the atoms of the structure or substructure, without any codification effort on the part of the user.

Structure search is considered to be a particular simpler case of substructure search. Figure 3 shows the natural language formulation of a query by substructure. Here, the question is: what are the compounds within the bank that contain this substructure and what data are related to these compounds? This is the most important type of query for a chemical data bank, and the

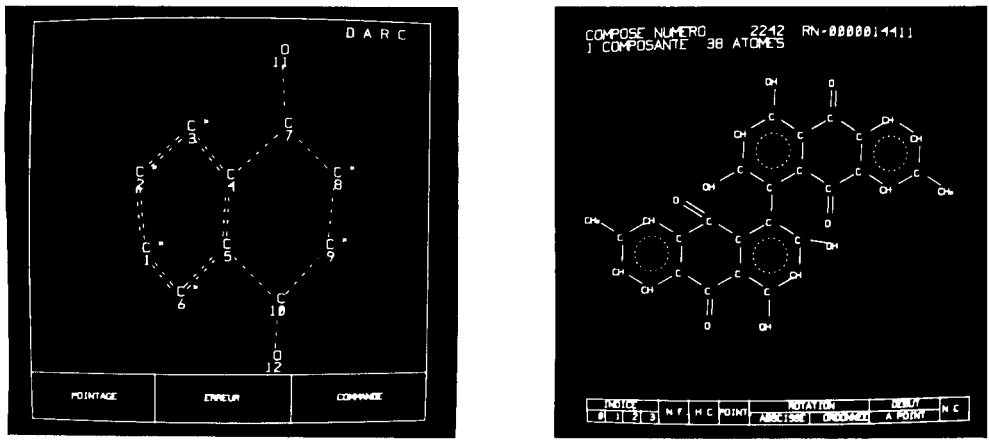


Fig. 3. Substructure searching based on structural formulae. The left-hand display shows the formulation of a query; the right-hand display is an example of a retrieved compound relevant to the query. The single dashed lines stand for "any bond". The stars on carbons 1, 2, 3, 6, 8, 9 indicate that these atoms can be substituted.

major interest of the DARC retrieval function lies in its processing. Indeed, this problem of substructure searching generally cannot be solved by the classical documentation software based on the nomenclature of the compound or by any other type of fragmentary encoding. Specific tools are needed to solve the problem efficiently at least in the case of an interactive retrieval system.

The substructure query is drawn on the screen. By definition, a substructure is not fully determined, because some atoms or bonds may remain undefined

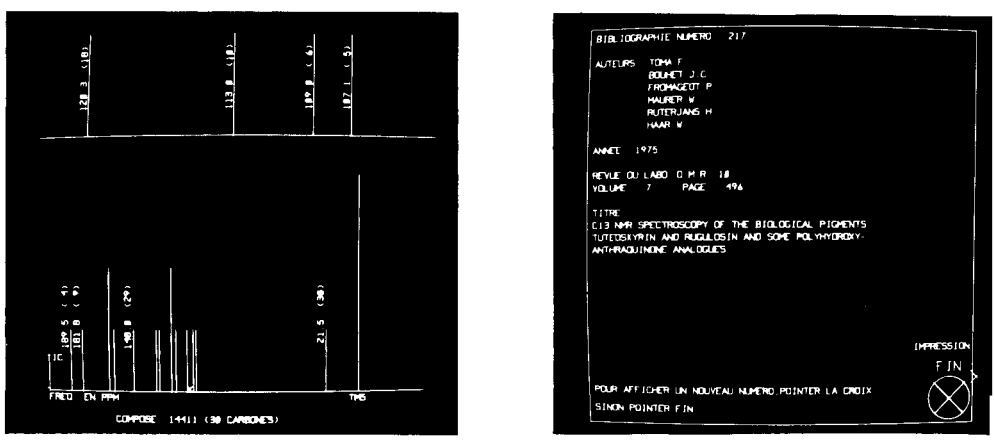


Fig. 4. Display of the ^{13}C-n.m.r. spectrum (left) and bibliographic data (right) related to the retrieved compound of Fig. 3. The upper part of the picture on the left was obtained by enlarging any part of the spectrum containing poorly separated peaks. Each peak bears its chemical shift in ppm. The number in parentheses refers to the corresponding carbon atom number.

or can be defined only by a list of possibilities. Moreover, the user should also be able to indicate the atoms for which substitution in the structure including the substructure is not forbidden. (These atoms are marked by a star.)

As an output to the substructure query, the user receives the structural diagram of each relevant compound selected by the search along with their related data. Figure 3 also provides an example of a relevant structure containing the sought substructure four times. The spectrum of this latter structure and its bibliographic source are given in Fig. 4.

Beyond the documentary use of the ^{13}C-n.m.r. data bank, its potential in

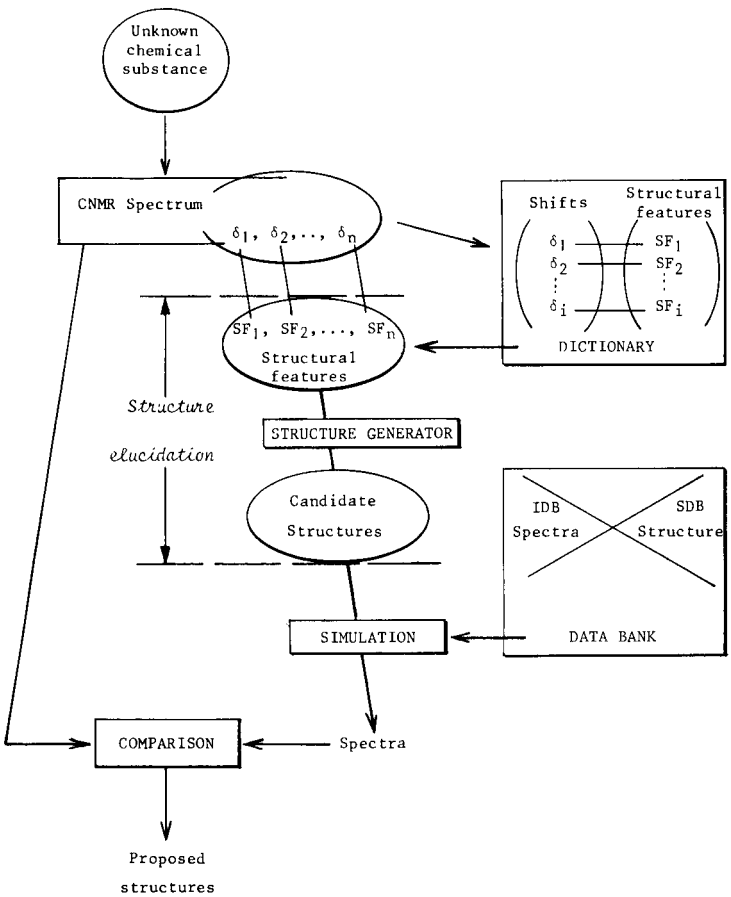


Fig. 5. General flow chart of the ^{13}C-n.m.r. structure elucidation procedure. When chemical shifts ($\delta_1,\ldots,\delta_n$) of a spectrum are sought in the dictionary, a list of structural features ($SF_1,\ldots,SF_n$) is obtained. These structural features are the input of a Structure Generator Program. The spectrum of the generated structures is simulated on the basis of the real data extracted from the data bank and compared with that requested. This comparison reduces the candidate structure list to the proposed structure list.

the area of computer-aided elucidation should be mentioned. The procedure for ^{13}C-n.m.r. elucidation is currently being implemented at the CIDA and surpasses the documentary aspect, by entering the field of artificial intelligence [6, 7]. It is obvious that the problem of elucidating an unknown structure on the basis of its n.m.r. spectrum is easily solved by consulting the bank when the spectrum itself or a very similar spectrum is in the bank, and is a mere retrieval problem. In this case, the aim of using an elucidation procedure is to propose a limited list of likely structures which are fairly reliable.

The general outline of the ^{13}C-n.m.r. elucidation procedure (Fig. 5) is as follows. Starting from the list of chemical shifts that constitute the spectrum, the consultation of a dictionary produces a list of structural features. These features are used as the input for a program which generates a set of candidate structures that satisfy these structural constraints. Moreover, the interrogation of the bank by means of the spectrum features and of the previous structural constraints leads to a list of real compounds and their related spectra. The candidate structures and the real data extracted from the bank make it possible for a correlation—prediction program to simulate the spectra of these candidate structures. Comparison of these simulated spectra with the requested spectrum allows a limited list of structures to be proposed for this spectrum.

It should be noted that this n.m.r. elucidation procedure is in fact designed to be used simultaneously with other spectral elucidation procedures, mainly those which operate on the mass spectra data bank.

REFERENCES

1 J. E. Dubois, in B. Dreyfus (Ed.), Proceedings of the Fifth Biennial International CODATA Conference, Pergamon Press, Oxford, 1977, p. 501.
2 J. E. Dubois, D. Laurent and H. Viellard, C.R. Acad. Sci., Paris, Ser. C, 263 (1966) 764.
3 J. E. Dubois, Research, 2 (1975) 32.
4 J. E. Dubois, in W. T. Wipke, S. Heller, R. Feldmann and E. Hyde (Eds.), Computer Representation and Manipulation of Chemical Information, Wiley, New York, 1974, p. 239.
5 J. E. Dubois, in F. Harary and A. T. Balaban (Eds.), Chemical Applications of Graph Theory, Academic Press, London, 1976, p. 333.
6 J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield and C. Djerassi, J. Am. Chem. Soc., 91 (1969) 2973.
7 J. E. Dubois, Israel J. Chem., 14 (1975) 17.

Analytica Chimica Acta, 112 (1979) 253—275
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

STRATEGIES IN MOLECULAR SPECTROSCOPIC ANALYSIS WITH APPLICATION OF QUEUEING THEORY AND DIGITAL SIMULATION

B. G. M. VANDEGINSTE

University of Nijmegen, Department of Analytical Chemistry, Toernooiveld, Nijmegen (The Netherlands)

(Received 23rd February 1979)

SUMMARY

Waiting lines in a laboratory for molecular spectroscopic analysis are simulated by digital methods. The influence of several parameters, such as the mean analysis time, the mean inter-arrival time of the samples and the routing procedure are evaluated by application of queueing theory. The queues were treated as a sample realization from a stationary first-order autoregressive system. The structure of the time series of the number of waiting samples in the model was identical to the time series observed in a real laboratory. The possibilities of digital simulation for decision-making in analytical laboratories are demonstrated.

Applications of operations research techniques in analytical chemistry have been reported. In particular, Massart et al. [1—4] have demonstrated that these mathematical techniques are suitable for the development of optimal strategies in chemical analysis. For example, Massart et al. [3, 4] applied graph theory successfully to optimize chromatographic separation schemes for multicomponent samples. Not only analytical procedures were optimized but also combinations or sequences of analyses; the optimal sequence of toxicological tests on a sample was established by dynamic programming in order to obtain a minimal mean delay time [1, 2]. In these studies, however, only properties of the analytical procedures were taken into account, and not the state of the laboratory itself, such as the capacity of the analytical procedures, the length of the queues and the number of available instruments and analysts.

In the work described here, a simulation model was constructed and validated for a molecular spectroscopic laboratory dealing with i.r., p.m.r., ^{13}C-n.m.r. and m.s. analyses. The time lag between the arrival of a sample and the production of analytical result is considered. Some relationships between variables of the laboratory are simulated, and the results of queueing theory are proved to be applicable in forecasting the effects of some variables on the delay time of the samples.

In many fields outside chemistry, digital simulation and queueing theory have often been applied to investigate the behavior of queueing systems, such as queues in airports, hospitals (waiting rooms, schedules of patients and

specialists), telephone networks, computer systems, etc. Despite these successful applications, adaptations for chemical purposes are few [5—8]. Väänänen et al. [7] simulated clinical laboratory activities comprising 64 different kinds of tests executed on 37 different types of equipment. The effects of an increased number of samples, and an increased number of analysts were calculated. Adeberg and Doerffel [8] applied some general results from queueing theory, but, to date, no applications of queueing theory or digital simulation to optimize activities in analytical laboratories have been reported.

The study of queues in analytical laboratories may be important. The delay of samples in an analytical laboratory normally exceeds the analysis time greatly. If a research department is considered with a sample traffic of λ samples per day to an analytical laboratory which receives the samples with a mean inter-arrival time $\overline{IAT}$ (the probability for $IAT < t$ is $p(IAT < t)$) and analyzes μ samples per day (mean analysis time $\overline{AT}$ and the probability for $AT < t$ is $p(AT < t)$) (Fig. 1), then it is clear that the profit (money, or research progress) of the research department will decrease with lower accuracy and lower probability ($p(T < t)$) of a delay (T) smaller than t. In contrast operation with smaller delays increases analysis costs, however, because of the larger facilities necessary to analyze the samples. A maximal net profit can be obtained when the profit and cost functions have the shapes shown in Fig. 2. As a consequence of the dead time of the research process itself, the profit of the research department will be relatively insensitive to the delay, if the delay is smaller than that dead time. However, for larger delays, profit may decrease rapidly, as the research progress is delayed. Anticipating the results from queueing theory, departments with a small workload will be relatively insensitive to alterations

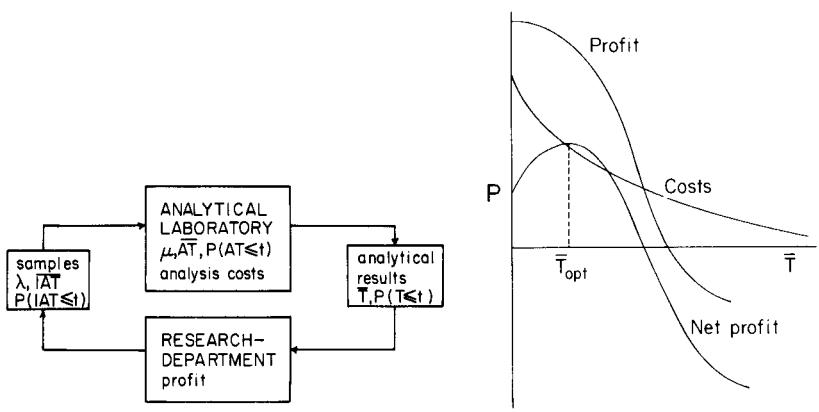


Fig. 1. The analytical laboratory in a closed loop with the research department (symbols are explained in the text).

Fig. 2. The analysis costs, research profit and net profit as a function of the average delay ($\overline{T}$) of the analytical result.

in the organization. Therefore, analysis costs will tend to increase exponentially with decreasing delays (lower workloads). For economic operation, it is essential to aim for this laboratory organization and strategy.

For a spectroscopic laboratory, the following questions should be answered.

Which strategy (or decision rules) should be followed to select the analytical method for solving the structure of the compound, given an estimated probability that several methods might solve the problem and the queue lengths in the laboratory? How to operate when an analytical method fails to elucidate the structure: after what time should the analysis be ended and another method tried? Which priority rules should be applied and how do they affect the waiting time? This question involves priority between samples from different research groups, priority between samples unsuccessfully analyzed, and priority for "easy" over "difficult" problems. What sort of priority should be applied? An absolute priority, which means that all samples of a higher priority are always analyzed before samples of a lower priority irrespective of their waiting time, or a priority depending on the waiting time? What is the effect of interruptions of the analytical procedure by other activities? What is the influence of the structure of the organization, e.g. a centralized versus decentralized organization, where each analyst can apply all analytical methods or only one method, respectively? Is it advantageous to accept samples only once a day?

Queueing theory answers the above questions only for simplified models such as single-server systems, with exponentially distributed inter-arrival times and analysis times of the samples (so-called $M/M/1$ systems) in a steady state. However, from calculations on theoretical $M/M/1$ systems, the effects of the variables mentioned may be estimated for larger systems.

In complex systems of networks of queues, such as analytical laboratories consisting of multiserver nodes and governed by state-dependent decision rules, the effects of these variables can only be calculated exactly from simulations with a model of the laboratory under consideration.

THEORY

The results obtained by queueing theory are derived for systems in a "steady state". That means that the arriving stream and service time are stochastic variables, which are described in terms of time-independent probability distribution functions. Consequently, many laboratories cannot be studied by queueing theory. For example, the sample input of some clinical and industrial control laboratories has a time-dependent probability distribution; in the early morning the laboratories are almost empty, and by the evening all samples have been completed. Jackson [9] demonstrated that for open networks of queues, where the arrival processes do not depend on the state of the system, each node can be considered individually. Because of the complexity of the network that represents analytical laboratories, and because of several unsolved theoretical problems in network queueing, exact calculations on analytical

laboratories by queueing theory have not been possible. However, queueing theory shows the important variables in queueing systems. From these theoretical considerations, the relative importance of various variables can be estimated for the construction of a simulation model of analytical laboratories.

Queueing theory: the behavior of M/M/*1 and related systems*

Queues are described by a shorthand notation $A/B/m$, where A,B and m represent the distributions for inter-arrival time, service time, and the number of channels. For example, in the $M/M/1$ system, both the inter-arrival time and service time are exponentially distributed, and there is only one service channel. That is the system for which most mathematical results are available.

The utilization factor (ρ) of the service channel plays an important role in all service systems. It is defined as:

$$\rho = \overline{AT}/(\overline{IAT}.m) \tag{1}$$

where $\overline{AT}$ is the mean analysis time, $\overline{IAT}$ the mean inter-arrival time and m the number of analysts serving the channel. This factor can be interpreted as the fraction of time the analyst is busy: $(1 - \rho)$ is the fraction that the analyst is idle, waiting for the next sample. Obviously, the system can only reach a steady state when $0 < \rho < 1$, because for $\rho > 1$, more samples arrive in the laboratory than can be analyzed, causing the number of waiting samples to grow in an unlimited fashion.

One of the first deduced relationships in queueing theory is the mean waiting time as a function of ρ for an $M/M/1$ system, where the samples are analyzed in the sequence of arrival at the system: first in—first out (FIFO) rule.

$$\overline{W} = \overline{AT^2}/2(1 - \rho).\overline{IAT} \tag{2}$$

where $\overline{AT^2}$ is the second moment of the analysis time. The asymptotic shape of the representation of eqn. (2) (Fig. 3) is very characteristic of all queueing systems. When this FIFO rule is applied in the laboratory, samples with large and small analysis times have an equal mean waiting time. Figure 3 shows that the mean waiting time depends strongly on the value of ρ. When $\rho > 0.85$, minor variations in the laboratory organization may induce a serious change in the waiting time. Therefore it is worthwhile to investigate in which nodes of the laboratory such a situation exists. In contrast, channels with overcapacity ($\rho < 0.5$) will be relatively insensitive to alterations in ρ as well as the organization.

When the analysis time is not exponentially distributed ($M/G/M$ systems), the mean waiting time depends linearly on the coefficient of variation $C_b^2 = \mathrm{VAR}(AT)/[E(AT)]^2$:

$$\overline{W} = \rho(1 + C_b^2)\,\overline{AT}/2(1 - \rho) \tag{3}$$

which is the well-known Pollaczek—Khinchin mean value formula.

By application of eqns. (2) and (3), the effect of decreasing the analysis time can be compared with the effect of automating or standardizing parts

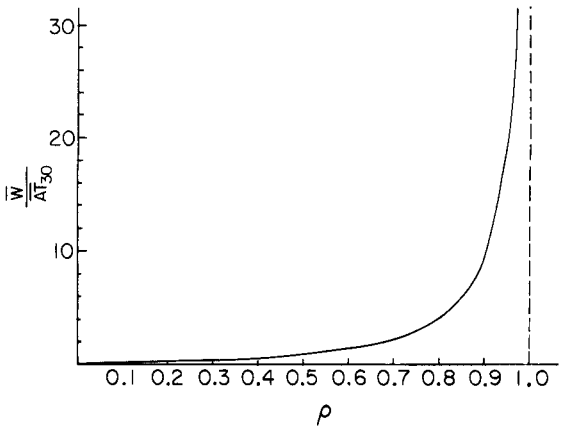


Fig. 3. The ratio between the average waiting time ($\overline{W}$) and the average analysis time ($\overline{AT}$) as a function of the utilization factor (ρ) for a system with exponentially distributed inter-arrival times and analysis times ($M/M/1$ system).

of the analytical procedure. For example, the analysis time can be reduced by modifying the analytical procedure or by presenting only single results rather than duplicates. In the latter case, the mean analysis time is halved, causing a reduction of the mean delay to 8% of its original value, when ρ decreases from 0.8 to 0.4 (Fig. 3) in an $M/M/1$ system. However, the probability of error is increased for single results, and the accuracy of faster analysis may be lower. Cost-profit analysis of the system (Fig. 2) should indicate the best mode of action. Automation and standardization of analytical procedures affects the coefficient of variation (C_b^2) of the analysis time. When this coefficient is reduced to zero, then the waiting time is reduced by a factor of 1.5 ($C_b^2 = 0.5$) to 3 ($C_b^2 = 2$) (eqn. 3), provided that the mean analysis time is not affected by the automation operation. These simple examples demonstrate that the statistical parameters of the analysis time have an important influence on the delay time. The practical conclusion is that alterations of analytical procedures in order to reduce waiting times should aim for a decrease of the mean analysis time, rather than for a decrease of the coefficient of variation for the analysis time.

In the particular case of structural analysis, analyses are not normally duplicated and standardization is difficult. Here the only way of modifying the parameters of the partition density function (p.d.f.) of the analysis time is to finish the interpretation of spectra after a certain time, AT_{max}, and to urge the analyst to pass the problem to another spectroscopic method. The original exponential p.d.f. of the analysis time takes the shape shown in Fig. 4 with first and second moments equal, respectively, to $\overline{AT}\,[1-\exp(-AT_{max}/\overline{AT})]$ and $2\overline{AT}\,[(\exp(-AT_{max}/\overline{AT}))\,(AT_{max}-\overline{AT})+\overline{AT}]$. When these terms are substituted in eqn. (2), the reduction of the waiting time is found as a func-

tion of the utilization factor (ρ) and the ratio between the maximal analysis time, AT_{max}, and mean, AT, of the original partition density function of the analysis time. Figure 5 demonstrates that truncating the analysis has the greatest effect at high ρ levels.

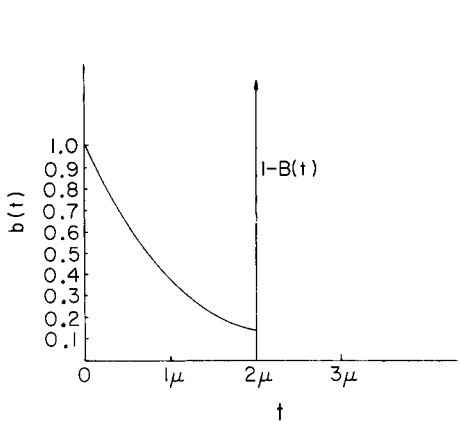
However, in the laboratory, two kinds of analyses are truncated: analyses which should be successfully finished if the analyst were allowed to study the spectrum for a longer time, and analyses which could not. Consequently, the truncation of analyses in one department may increase the flow towards the other departments, and therefore increase the overall waiting time. It depends on the utilization factors of the other departments, whether the performance of the total system is improved or not when analyses are interrupted. Generally speaking, it is possible to circumvent bottleneck situations in one particular node in this way, if the utilization factors of the other nodes are relatively low.

Queueing theory: interruptions of the analytical process for other activities

The mean lengths of the busy, E(busy), and idle, E(idle) periods of the channel (or analyst) are two important features of queueing systems. The idle time is defined as the time during which no samples are in the channel. For all systems with an exponentially distributed inter-arrival time, the mean idle period equals $E(IAT)$, which is independent of the utilization factor. From the definition of ρ, it follows that:

$$\rho = E(AT)/E(IAT) = E(\text{busy})/(E(\text{idle}) + E(\text{busy})) \tag{4}$$

$$\text{or } E(\text{busy}) = \rho/(1-\rho)\, E(IAT) \tag{5}$$



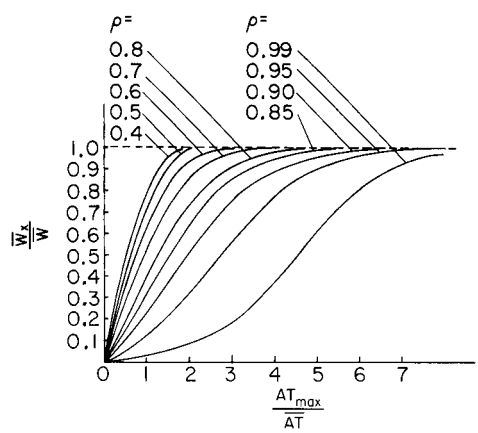


Fig. 4. The truncated analysis time p.d.f. $b(t) = \exp(-\mu t)$ for $t < 2$, and $b(t) = \exp(-2\mu)$ for $t > 2$.

Fig. 5. Reduction of the average waiting time, as a function of the maximal allowed analysis time, AT (max).

This means that for $\rho = 0.9$, $E(\text{busy}) = 9 \times E(IAT)$. Therefore, the mean busy period of an analyst receiving 1 sample per hour, on average equals 9 h. Therefore obviously in laboratory practice, it may happen that no analyses are done, although samples are waiting, i.e. the service is interrupted for other activities.

The question arises how the mean interruption time (β) and the mean time between the interruptions (α) affect the mean waiting time. Furthermore, one may desire to compare the case where the "on" and "off" times are exponentially distributed, with the case they are constant or scheduled. The condition for a steady state is that $\beta/(\alpha + \beta) \leqslant 1 - \rho$. The effect depends on how the interruptions are scheduled. Two situations can be distinguished.

(i) Other activities are permitted as soon as there are no samples in the channel, and the arrival of a sample does not end these activities prematurely. The analyst immediately starts a new period of other activities when he finds the system empty at the end of other activities. The mean waiting time is increased by β for exponentially distributed interruptions [10]. For constant durations of all interruptions, the mean waiting time is increased by $\beta/2$.

(ii) Other activities may also be started during a busy period. The factor (R) with which the mean waiting time of the system is increased when interruptions of the analysis are permitted, depends on the p.d.f. of α and β. According to Fisher [11] when α and β are exponentially distributed,

$$R = \{\overline{AT} + [(\alpha\beta)^2/(\alpha + \beta)^3\rho][1 - \rho]\}/\{[\alpha/(\alpha + \beta) - \rho]\,\overline{AT}\} \tag{6}$$

For α and β constant

$$R = (1 - \rho)/[\alpha/(\alpha + \beta) - \rho] \tag{7}$$

From the plots of eqns. (6) and (7) in Fig. 6, the influence of the interruptions of the analyses on the mean waiting time is demonstrated to become important

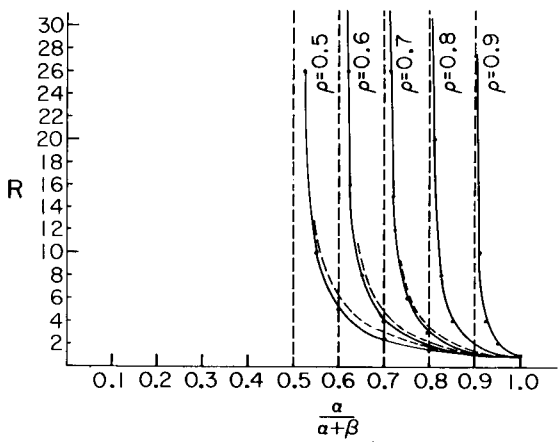


Fig. 6. The ratio (R) between the average delay time for interrupted and uninterrupted analyses, as a function of the available time for analysis. (—) α and β exponentially distributed; (---) α and β constants.

as $\lim R = \infty$ for $\alpha/(\alpha + \beta) \to \rho$. In fact, interruptions cause a system with a low utilization factor to behave as a system with a high one. That means that the waiting time becomes asymptotically dependent on ρ, even for small values of ρ. Furthermore, Fig. 6 indicates that the p.d.f. of the interruptions has little effect on the waiting time. Thus it is only slightly better to schedule interruptions, e.g. at the end of the day for administration, or getting chemicals, etc. However, other activities scheduled only during the idle period of the queueing system have a negligible effect on the mean waiting time. Obviously, eqn. (6) can be used to calculate the effect of equipment failures.

Priorities in queueing theory

All equations in the above paragraphs are valid for FIFO disciplines. Of course, there may be many reasons to deviate from this FIFO rule, some samples being given priority. In an analytical laboratory, priority can be given to samples depending on their origin (different research groups), their analysis time ("easy" and "difficult"), or their history in the laboratory (first, second analytical method which is tried). A priority difference can be obtained by attributing urgency numbers (U_p) to the samples, which may be a function of the time that a sample waits in the system (W_p), and an urgency parameter (b_p), which is characteristic for the priority group p:

$$U_p = W_p^r \cdot b_p \tag{8}$$

The samples are then analyzed in the sequence of decreasing urgency numbers. For a large value of r, the sequence becomes FIFO, while for small values, an absolute priority discipline is obtained, because the urgency numbers are equal to b. Kleinrock [12] derived the mean waiting time ($\overline{W}_p$) of the pth priority group in an $M/M/1$ system with P priority groups, having time-dependent urgency numbers from eqn. (8):

$$\overline{W}_p = \left\{ [W_0/(1-\rho)] - \sum_{i=1}^{p-1} \rho_i \overline{W}_i \, [1 - (b_i/b_p)^{1/r}] \right\} \Big/ \left\{ 1 - \sum_{i=p+1}^{P} \rho_i [1 - (b_p/b_i)^{1/r}] \right\} \tag{9}$$

where $W_0 = \sum_{i=1}^{P} \rho_i \, . \, \overline{AT}_i$, $b_p < b_{p+1}$, $p = 1, 2 \ldots . P$.

Substituting $r = 0$, the mean waiting time is found for the samples of the pth priority group with absolute priority over all samples of group $1 \ldots (p-1)$:

$$\overline{W}_p = \left\{ W_0/(1-\rho) - \sum_{i=1}^{p-1} \rho_i \, . \, \overline{W}_i \right\} \Big/ \left\{ 1 - \sum_{i=p+1}^{P} \rho_i \right\} \tag{10}$$

where $p = 1, 2, \ldots . P$. From the plot of eqn. (9) in Fig. 7 it is easily seen that by choosing appropriate values of r and b, each ratio between the mean waiting times of the various classes can be obtained. Moreover, these equations show that when the various priority groups have equal analysis times, the mean waiting time of the whole population of samples is not influenced by any priority discipline.

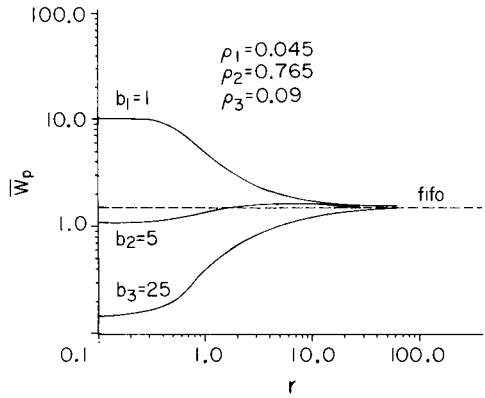


Fig. 7. Time-dependent priorities, varying from absolute priority (r = 0.1) to FIFO discipline (r = 100); b_1: b_2: b_3 = 1:5:25 (eqn. 9).

Accordingly, optimization of b and r is not achieved by minimizing the mean waiting time exclusively, but should include cost functions for waiting [13] for the different priority groups of samples. Otherwise, for priority groups with different mean analysis times, the overall lowest mean waiting time is found by giving absolute priority to the samples with the shortest analysis time. Figure 8 shows the effect of this discipline as a function of the ratio of analysis times, and number of samples in both classes.

In the analytical laboratory, this situation occurs when an analyst does two different analyses, or when samples can be subdivided into two groups, e.g., "easy" and "difficult" samples with "small" and "large" analysis times, respectively, i.e. the interpretation of a spectrum may be easy or difficult. In particular, many examples [14] indicate that this separation into two groups provides a considerable reduction in mean waiting time, as opposed to the FIFO system.

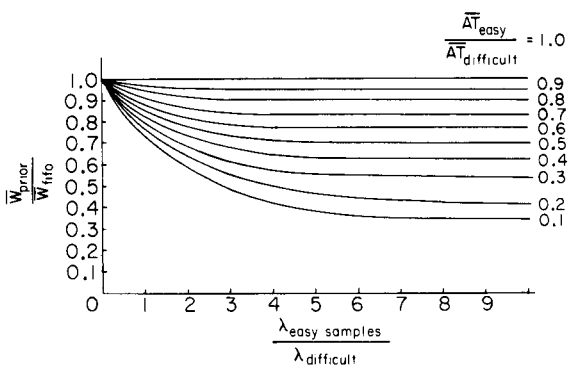


Fig. 8. Reduction of the average waiting time, attributing absolute priority to the "easy" samples as a function of the ratio of the analysis time and ratio of the input density of "easy" and "difficult" samples.

In many cases, however, it is not possible to determine previously if a spectrum will be easy or difficult to interpret. Thus, after a certain interpretation time (q), the difficult spectrum may be transferred to a pile of unfinished spectra, and the measurement of the next sample or interpretation of the next spectrum is started.

This is a common method of job handling in time-shared computer systems, and is known as round-robin scheduling. Kleinrock [12] demonstrated that round-robin scheduling has no effect on the mean waiting time of the total population of samples, but shortens the mean waiting time of the easy samples, at the expense of the difficult ones. For q approaching zero, the waiting time of a sample becomes linearly proportional to its analysis time, and a discrimination is obtained in favor of short analyses.

Subdividing the interpretation time of a spectrum into infinitesimally small steps is unrealistic. Therefore the effect of applying round-robin scheduling was simulated here only during the interpretation step, and for large values of q. Figure 9 (a, b) shows the results of simulations of a laboratory system where the interpretation time for difficult samples is twice that for easy samples. For a negligibly small measurement time, "easy samples" wait for shorter times only when the interpretation is done in steps smaller than 0.1th of the mean interpretation time. Moreover, the effect is considerably decreased when the measurement and interpretation time are about equal, and only the interpretation step is partitioned (Fig. 9b).

These simulations lead to the conclusion that in spectroscopic analysis,

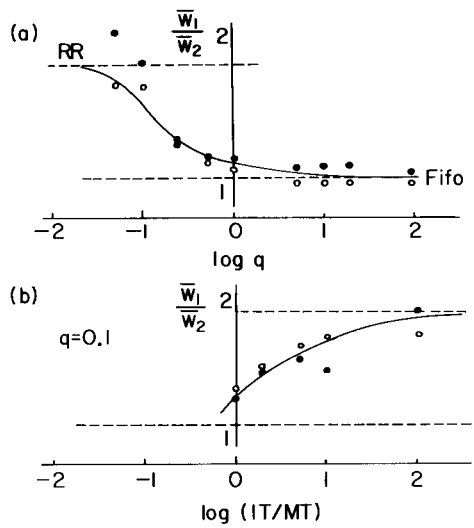


Fig. 9. (a) The effect of applying round-robin scheduling as a function of the ratio of the step length and the mean interpretation time (q = step length/$\overline{IT}$) when the variation coefficient (C_b^2) of the analysis time is 1 (○) and 3 (●). (b) The effect of applying round-robin scheduling with q = 0.1 as a function of the ratio of the interpretation time (IT) and the measuring time (MT) of the sample.

waiting time can be improved considerably only by giving priority to the "easy" samples which have previously been recognized as such.

Description of the model

The laboratory model basically consists of four spectroscopic sections (i.r., p.m.r., ^{13}C-n.m.r. and mass spectrometry) for structural analysis with the sample flow (samples per day) shown in Fig. 10. In the terminology of queueing theory, these four sections form an open network, i.e. a network of queues in which samples originate from external sources, and each sample eventually leaves the system. Under equilibrium conditions the total external input flow rate to the network must equal the total external departure flow rate.

Parameter estimation of the model

For simulation of the model, the partition density functions of the arrivals of the samples, the measurement time and interpretation time of the analytical procedures were measured in an actual laboratory situation, and their parameters were then calculated. The memoryless property of the input process is proved from calculations of the autocorrelation functions of the input density of the total laboratory and the separate sections. As an example, the autocorrelation function of the total sample rate is shown in Fig. 11. The output of the laboratory is characterized by the mean and standard deviation of the delays for the different types of sample, dependent on history, procedure, etc. (Table 1).

An important variable is the number of waiting samples at the various sections of the laboratory. Figure 12 shows the fluctuations in the number of

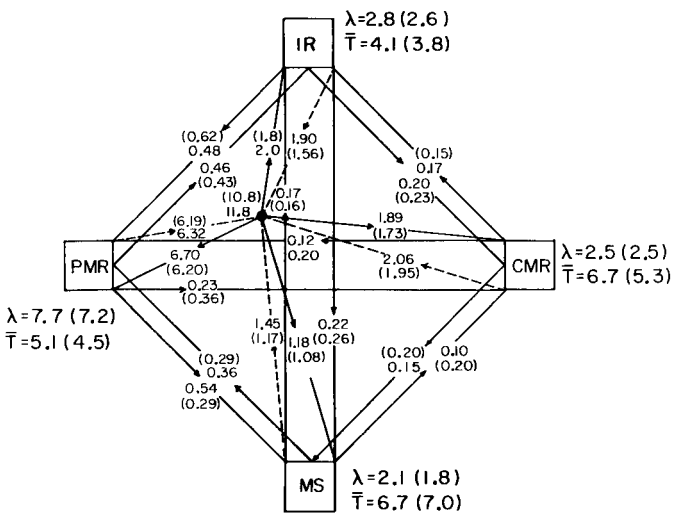


Fig. 10. The sample flow (samples per day) through the laboratory network (numbers in parentheses are simulated data).

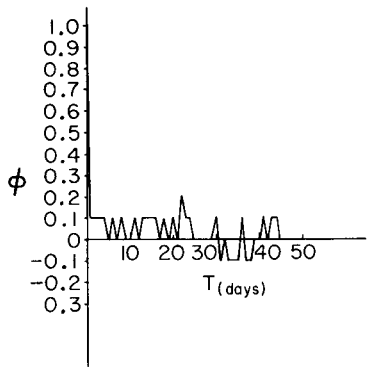


Fig. 11. The autocorrelation function (Φ) of the input density to the laboratory.

TABLE 1

Mean delay time ($\overline{DT}$), variance[a] of the mean delay time [VAR($\overline{DT}$)], autocorrelation value [Φ_1] at $\tau = 1$, and Student's t value between actual and simulated data

Section	Actual data			Simulated data			t[b]
	$\overline{DT}$	VAR($\overline{DT}$)	Φ_1	$\overline{DT}$	VAR($\overline{DT}$)	Φ_1	
I.r.	4.1	0.05	0.17	3.8	0.1	0.81	0.8
P.m.r.	5.1	0.02	0.27	4.5	0.03	0.81	2.5
M.s.	6.7	0.11	0.09	7.0	0.8	0.86	0.2
^{13}C-n.m.r.	6.7	0.09	0.19	5.3	0.03	0.46	4.3
Samples solved by the same analytical sections							
I.r.	6.2	0.21	0.19	6.4	0.15	0.32	0.3
P.m.r.	5.9	0.03	0.24	6.4	0.03	0.51	2.0
M.s.	9.8	0.39	0.18	11.2	0.75	0.45	1.3
^{13}C-n.m.r.	8.3	0.17	0.11	8.6	0.09	0.09	0.6
Samples which visited several sections							
1	5.1	0.01	0.32	5.4	0.03	0.80	1.9
2	11.1	0.19	0.15	11.5	0.20	0.56	0.5
3	19.7	9.16	−0.07	18.8	3.75	0.77	0.4
4	26.8	9.16	0.04	30.2	4.51	0.69	0.9

[a] Variance corrected for autocorrelation in the data.
[b] $t_{0.005}(250) = 2.6$.

samples in the queue for each analytical method in the laboratory. A sample interval of one day was used. In order to obtain a model for this network of queues, the four queues were treated as a sample realization from a stationary first-order autoregressive system AR(1) model, which has the form

$$n_t = \Phi_1 . n_{t-1} + a_t \tag{11}$$

where n_t is the difference in the level from the mean level of the number of

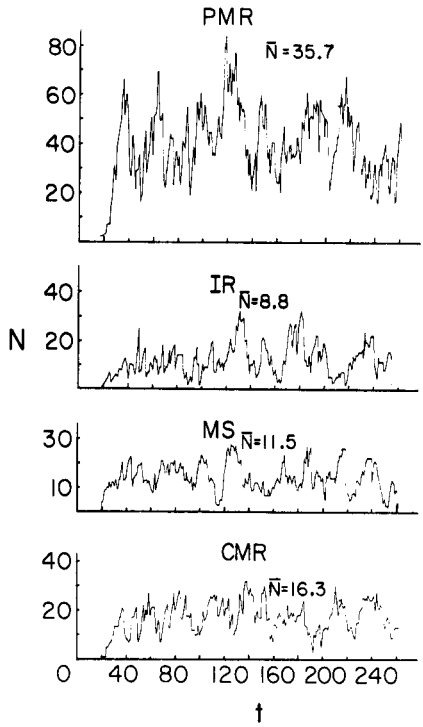


Fig. 12. The time series of the number of samples waiting at the four sections of the laboratory.

waiting samples at time t ($n_t = N_t - \overline{N}$); n_{t-1} is the difference in level from the mean level of the number of waiting samples at time $t - 1$; Φ_1 is the autocorrelation coefficient at $\tau = 1$; and a_t is a random variable which is normally and independently distributed with zero mean and variance σ_a^2, called the residual variance.

When the AR(1) model is adequate, it implies that the dependence of n_t on its past history is completely accounted for by the term $\Phi_1 n_{t-1}$ in the model. For an AR(1) model, the residual variance is $\sigma_a^2 = \sigma_N^2 (1 - \Phi_1^2)$, where σ_N^2 is the variance of the number of waiting samples. From the calculation of the autocorrelation function, the parameters Φ_1 and σ_a^2 for each queue were calculated (Table 2). The essential validation of the simulation model will consist of the inference that the structure of the observed time series and the structure of the simulated time series are identical for the simulation of the actual situation.

Structure of the simulation model: fixed characteristics

Sample flows are generated at three different origins. The model is run for 250 days (4000 samples) and the resulting statistics are examined. Various

TABLE 2

Mean, autocorrelation value at $\tau = 1$ (Φ_1), variance of number of samples in the analytical sections and variance[a] of the mean of actual and simulated data. Student's t values[b] between means and $G(0,1)$ test[c] on Φ_1 and residual variance

Section	Actual data				Simulated data					
	$\overline{N}$	VAR(N)	VAR($\overline{N}$)	Φ_1	$\overline{N}$	VAR(N)	VAR($\overline{N}$)	Φ_1	t	$G($(
I.r.	11.5	46.2	1.94	0.83	8.8	27.4	0.76	0.79	1.6	3.7
P.m.r.	39.1	228.0	8.3	0.81	35.7	444.8	42.5	0.92	0.5	1.9
M.s.	13.9	34.8	1.6	0.85	11.5	69.9	8.9	0.90	1.0	2.7
^{13}C-n.m.r.	16.8	43.6	1.6	0.81	16.3	57.4	2.9	0.89	0.2	2.2
Lab	81.4	582.7	24.4	0.83	72.3	789.6	77.3	0.92	0.9	2.2

[a]Variance corrected for autocorrelation in the data.
[b]$t_{0.005}(250) = 2.6$.
[c]$1/2\ \chi^2_{0.01}(3) = 5.7$.

partition density functions of the inter-arrival times and lengths for the series of the three sources of samples can be chosen independently from each other as indicated below.

Partition density function	Mean	Variance
Exponential:		
$f(x) = \lambda \exp(-\lambda x) \quad (x > 0)$	$\overline{X} = 1/\lambda$	$1/\lambda^2$
K-Erlangian:		
$f(x) = r\lambda(r\lambda x)^{r-1} \exp(-r\lambda x)/(r-1)!$	$\overline{X} = 1/\lambda$	$1/(r\lambda^2)$
Hyperexponential:		
$f(x) = \sum_{i=1}^{R} \alpha_i \lambda_i \exp(-\lambda_i x)$ $\sum_{i=1}^{R} \alpha_i = 1$	$\overline{X} = \sum_{i=1}^{R} \alpha_i/\mu_i$	$2\sum_{i=1}^{R} (\alpha_i/\mu_i^2) - \left[\sum_{i=1}^{R} \alpha_i/\mu_i\right]^2$
General:		
$f(x)$ = arbitrary	—	—
Deterministic:		
$f(x)$ = constant	—	0
Gaussian:		
$f(x) = (1/\sigma\sqrt{2\pi}) \exp[-\frac{1}{2}((x-\mu)/\sigma)^2]$	μ	σ^2
Poisson:		
$P_k(t) = [(\lambda t)^k/k!] \exp(-\lambda t)$	λt	λt

For each sample arriving at the laboratory, the probabilities of a successful analysis of the molecular structure are estimated for the various analytical procedures. These probabilities are estimated on three levels: $p = 0$ indicates that the analytical procedure cannot furnish the structure; $p = 0.5$ and $p = 1$. These estimations were assumed to be exact. The fraction of samples having $p = 0$—1.0 for each analytical procedure is determined from the observed sample flow in the laboratory. The comparison of the simulated sample flow (numbers in parentheses in Fig. 10) and the actual data demonstrate that these estimates reproduced the actual situation very well. The interpretation and measurement times of the samples were generated by taking random numbers from a generalized exponential or Erlangian distribution, depending on the desired value of the variation coefficient.

The laboratory has a maximum capacity of 20 analysts and 10 instruments. Each analyst has a fixed experience with the four analytical procedures, expressed in a parameter between one (fully qualified) and zero (no experience). This factor combined with a minimal required experience for a particular analytical procedure changes the organization of the laboratory from an open organization, where all analysts can do all analyses, to a closed organization, where the analysts specialize in one analytical method only. Furthermore the analysis time becomes a function of the experience parameter. Priority factors (between 1 and 0) are assigned to the various sources of samples. Three kinds of interruptions of the analyses are distinguished: other activities, coffee breaks and holidays.

Structure of the simulation model: variable operating decisions

The strategies considered in the model are as follows.

(A) Strategies concerning sample priorities. (1) The sample in queue with the earliest laboratory arrival date is selected first. (2) The sample in queue with the earliest arrival date at the spectroscopic unit is analyzed first. (3) Samples in each section are subdivided into classes according to the number of analyses unsuccessfully done before. Priority (varying from FIFO to absolute priority) is assigned either to samples which have visited the largest number of sections, or to samples which have visited the smallest number of sections. (4) Samples receive priority according to the analysis time expected: the shortest-expected-analysis-time first (SEAT) discipline.

(B) Strategies based on the choice of analytical method (routing strategies). (1) The analytical method with the highest probability of a successful eludication of the requested structure is selected. This is a fixed routing procedure in which a sample path is uniquely determined from the properties of the sample itself. (2) Not only the probability of success but the expected waiting times for the various methods are considered when the sample enters the laboratory. The relative importance of these criteria is given by weighting factors. Because this routing algorithm bases its decisions on the observed traffic flow, it is called a dynamic or adaptive routing procedure.

(C) Strategies for assignment of analysts. An analyst may only be assigned

to a section which has an unmanned instrument. There are two main possibilities: (1) complete centralization — when an analyst completes an analysis, he is available for reassignment to another method; (2) complete decentralization — the analyst is always assigned to the same method. In the centralized situation, the assignment of analysts can be governed by the following work selection rules: (i) assign available analysts to the method with which they have most experience; (ii) assign available analysts to the method with the oldest sample in its queue and for which the analyst has sufficient experience; (iii) the experience of the analyst and the waiting time of the oldest sample in each queue are weighted.

(D) Stopping rules for the analysis. (1) Analyses are not truncated. (2) A maximal measuring and interpretation time is assigned to each analytical method, regardless of the originally estimated probability of success and the number of unsuccessfully applied methods. (3) The maximal measuring and interpretation time is a function of the probability of success and the number of methods tried. (4) The maximal measuring and interpretation time is also a function of the state of the laboratory, i.e. the number of samples waiting for each analytical method.

Validation of the simulator

The validation of a simulation model requires a comparison of actual versus simulated data. It is common to compare histograms of both data series, employing the standard χ^2, t and F statistics. However, in most cases, the actual data as well as data from many simulations are not independent but serially correlated. This greatly complicates the application of the above statistics. Moreover, serial correlation in time itself is often an important characteristic of the system simulated. Hsu and Hunter [15] suggested comparison of actual and simulated data by identifying a time series model and estimating the parameters of this model by the techniques outlined by Box and Jenkins [16]. Then the models are tested for differences in their means, autoregressive parameters and residual variances. The proposed simulation model was validated by two methods.

(1) Comparison of the means of the delay times of different groups of samples. This involved: (a) mean delay time at the different sections (methods); (b) mean delay time of all samples with the same final analytical method; (c) mean delay time of all samples with the same number of analyses.

This comparison was done by applying Student's t test for correlated time series, as described by Box and Tiao [17]. Therefore, first of all, the correlated AR(1) time series N_t was transformed to independent normal variables (u_t) with mean $\bar{u}$ and variance s_u^2, by applying the equation

$$u_t = (N_t - \Phi_1 N_{t-1})/(1 - \Phi_1) \tag{12}$$

The variance (s_u^2) of a transformed AR(1) time series, then equals [15]

$$s_u^2 = s_N^2 (1 + \Phi_1)/(1 - \Phi_1) \tag{13}$$

and the mean $\bar{u} = \bar{N}$. Thereafter, Student's t test between the actual (N_t) and simulated data (N_t^*), which were respectively transformed to the independent normal variables u_t and u_t^*, with means $\bar{u}$ and $\bar{u}^*$ and standard deviation s_u and s_u^*, was applied according to the equation

$$t = (\bar{u} - \bar{u}^*)/[s_u^2/(n-p) + s_u^{*2}/(m-p)]^{\frac{1}{2}} \tag{14}$$

where n and m are the number of observations in each series.

(2) Comparison of the autoregressive parameters and residual variances of the number of samples in the queues at the end of each day. For two first-order autoregressive time series, the inferences that $\psi = \Phi_1 - \Phi_1^* = 0$, and $\gamma = \sigma_a^2/\sigma_a^{*2} = 1$ should be tested. This can be done by using an inferential statistic $G(\psi, \gamma)$, distributed as a χ^2 variable. In the first stage, the test is made to establish whether the inference $\psi = 0$, $\gamma = 1$ is tenable or not. If it is not, then the test is continued to establish if this is due to ψ or γ or both. Because the expressions of the G statistic are fairly complicated, the outline given by Hsu and Hunter [15] should be consulted.

EXPERIMENTAL

For the data collection, during the year 1977, the arrival rates of all samples at the different analytical sections, and the departure of all analytical results were recorded, without additional load on the laboratory personnel. The interpretation and measurement times of about 200 samples were timed.

Because the duration of other activities were very difficult to assess without affecting the laboratory activities, the magnitudes of these parameters were estimated from interviews with the laboratory technicians.

The simulation program was written in FORTRAN IV, including the GASP [18] subroutine package. A simulation run for one year of laboratory operation takes about 10 min of CPU time in an IBM 370/158 computer system.

RESULTS AND DISCUSSION

This discussion will be restricted to the statistical comparison of the output of the proposed computer model with the actual situation in the laboratory. A further paper will elaborate on the effect of variations in the decision rules. Comparison will be made between simulated results and the effects predicted by queueing theory.

For the simulation of the actual situation, the variable operating decision rules of the laboratory were determined. Interviews with the laboratory personnel indicated that strategies B(1), C(2) and D(1) were applied. The strategy concerning the sample priorities, based on the history of the sample, was elucidated from the collected data. In almost all sections, samples which had been analyzed before had to wait longer for further analysis than unanalyzed samples, i.e., the two groups of samples were given different priorities (Table 3).

TABLE 3

Mean delay time, variance[a] of mean delay time, autocorrelation value at $\tau = 1$, and Student's t values between actual and simulated data of samples which were and were not analyzed before in another section

	Actual data			Simulated data			t^b
	$\overline{DT}$	VAR($\overline{DT}$)	Φ_1	$\overline{DT}$	VAR($\overline{DT}$)	Φ_1	
Not analyzed before							
I.r.	3.9	0.07	0.40	2.8	0.66	0.40	3.6
P.m.r.	5.1	0.02	0.32	4.5	0.87	0.32	1.9
M.s.	5.8	0.13	0.22	5.0	0.78	0.22	1.2
^{13}C-n.m.r.	6.1	0.10	0.27	5.6	0.75	0.27	1.2
Lab	6.1	0.10	0.27	4.4	0.36	0.27	5.1
Analyzed before							
I.r.	4.7	0.25	0.03	5.1	0.57	0.03	0.6
P.m.r.	5.3	0.20	0.08	7.5	0.69	0.08	3.5
M.s.	7.9	0.32	0.	8.5	0.59	0.09	0.6
^{13}C-n.m.r.	8.9	0.48	0.03	9.0	0.39	0.03	0.2
Lab	6.5	0.08	0.06	7.5	0.26	0.06	3.1
Student's t *values between mean delay of samples which were and were not analyzed before in another analytical section*							
I.r.		0.45			4.8		
P.m.r.		1.46			6.0		
M.s.		3.5			3.9		
^{13}C-n.m.r.		3.6			6.4		
Lab		4.5			16.2		

[a]Variance corrected for autocorrelation in the data.
[b]$t_{0.005}(200) = 2.6$.

A comparison of the ratio of the delay times for samples subjected to 1, 2, 3 and 4 procedures in the laboratory (Table 1), and those found for various priority rules in the model, indicated that only a dynamic priority discipline between n and $(n + 1)$ times analyzed samples could give the same results for the model and the actual situation.

Such a priority discipline assigns fixed priority numbers (P) to the samples arriving at a queue. The sample with the smallest sum of priority number and arrival date at the queue is analyzed next. This discipline has the effect of considering some samples to have a higher priority than others, but takes into account the undesirability of leaving samples with low priority for too long, as occurs in the absolute priority case.

The model assigned priority numbers (P) to the samples as a function of the number (N) of sections the samples have visited. Therefore that number of visits was multiplied by an urgency number (A) that defines the priority difference obtained between the various categories. $P = A*N$. By varying A from −100 to +100, the discipline rule is altered from giving absolute priority

to the samples which visited the smallest number of sections, to the reverse situation. For $A = 0$, samples are analyzed in the sequence of arrival at the analytical section, and a simulation with $A = -20$ gave the same results as priority rule A(1) (see above). Figure 13 illustrates the mean delay times for samples subjected to 1, 2, 3, and 4 procedures as a function of A. As shown by eqn. (9), the overall mean waiting time of all samples is not influenced by any priority rule, provided that the mean analysis times of each priority group are equal. Furthermore, Fig. 13 shows that the laboratory applies a priority rule that has the same effect as a dynamic priority discipline with $A = -20$. However, from Fig. 13 more general conclusions can also be drawn. For $A = -100$, the differences between the mean delay times of the four types of samples are much greater than for $A = +100$, while the overall mean delay time remains unaffected. As a consequence, the variance of the delay time and the longest delay time found in the laboratory will be a function of A, as shown by Fig. 14. Here a minimal value for both performance variables is found for $A = +100$.

From this experiment with the model, the laboratory was recommended to change their priority rule in order to decrease the variances of overall delay time. Future observations of the laboratory should indicate the validity of the forecast.

Verification of some properties of network queues

Little's result. One of the important relations in queueing theory is known as Little's result, which relates the average number of samples in the system ($\overline{N}$) to the average arrival rate (λ) and the average time ($\overline{T}$) spent in the system: $\overline{N} = \lambda \overline{T}$. This relation is independent of the distributions of the input flow and analysis time, and is therefore valid for all kinds of queueing systems. Furthermore, the relation depends neither on the number of analysts in the system nor on the particular queueing discipline. From Tables 1 and 2, it is easily verified that this relation indeed holds for each section of the laboratory under study.

Mean delay time. The total flow to each node in the network is the sum of the external flow (λ_i) to the node and the flow coming from the other nodes. According to Lemoine [19], the mean delay time for all samples can be calculated from the average number of samples in the nodes ($\overline{N}_i$) and the total external flow (λ).

$$\overline{T} = \sum_{i=1}^{M} \overline{N}_i \Big/ \sum_{i=1}^{M} \lambda_i = \left(\sum_{i=1}^{M} \overline{T}_i \gamma_i \right) \Big/ \lambda \tag{15}$$

This equation was used to calculate a mean delay time in the laboratory of 6.92 days, which agrees reasonably well with the observed mean delay of 6.88 days. Jackson [9] demonstrated that, for such systems, the number of samples at each time point should be independent random variables. The calculated cross-correlograms between the number of samples at the various nodes proved this property of the laboratory, as no correlation coefficients significantly differing from zero were found.

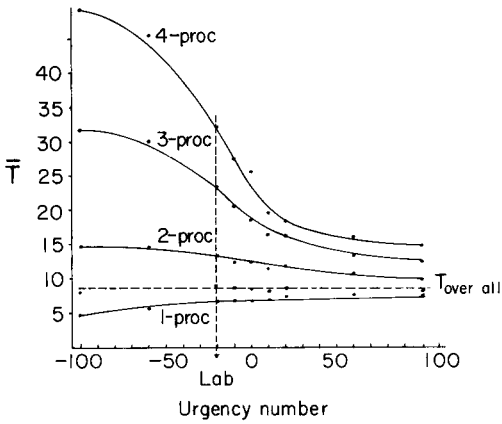


Fig. 13. The average delay time of samples which visited n sections as a function of the urgency number (A) defining the priority difference between samples that visited N and $(N + 1)$ departments.

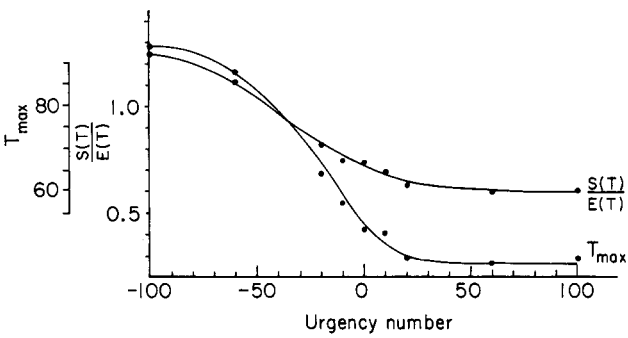


Fig. 14. The maximal delay time, T(max), and variation coefficient of the delay time as a function of the urgency number (A), defining the priority difference between samples that visited N and $(N + 1)$ departments.

The balance or conservation equation of the network is:

$$\gamma_i = \lambda_i + \sum_{j=1}^{M} \gamma_j \, p(i,j) \tag{16}$$

which says that the equilibrium rate of flow through node $i(\gamma_i)$ is the sum of the external input rate λ_i and the total rate of internal transfers to node i, with $p(i, j)$ the probability that a sample from node j is transferred to node i. Table 4 shows these probabilities calculated for the laboratory investigated.

Validation

Queue lengths. In the model proposed, the queue length at an analytical

TABLE 4

Transfer probabilities

I/J	P.m.r.	I.r.	M.s.	^{13}C-n.m.r.
P.m.r.	—	0.17	0.17	0.05
I.r.	0.06	—	0.08	0.07
M.s.	0.07	0.08	—	0.06
^{13}C-n.m.r.	0.03	0.07	0.03	—
Out	0.83	0.68	0.70	0.83

section is defined as the sum of all waiting and unfinished samples, i.e. the sum of samples and uninterpreted spectra. Student's t tests (eqn. 14) could not detect any significant difference between the mean queue lengths in the analytical sections or between the total mean queue length of model and laboratory (Table 2).

Comparison of the autoregressive parameters Φ_1 and residual variances of the number of samples in the queues at the end of each day, in model and laboratory, by means of G-statistics (Table 2) demonstrated that the dynamic structure of the simulated queue length series is indistinguishable from that of the observed series. Both series are adequately described by an AR(1) model, with equal parameters. This means that no differences in magnitude and velocity of queue fluctuations in model and laboratory could be detected.

Delay times (DT). Various delay times of samples can be distinguished in the model and in the laboratory. In the first place, the mean delay time in each analytical section (Table 1) is considered. Here, samples which were or were not analyzed before in another section can be distinguished (Table 3). Secondly, the mean delay time of the group of samples, for which the structure can be solved by the same analytical method, is considered. Finally, the mean delay of all samples which visited the same number of sections is compared. Again, Student's t tests could not detect significantly differing mean delay times between model and laboratory. Table 3 also proves that the strategy concerning sample priorities, as applied in the model, gives rise to longer delay times at each analytical method for samples which were subjected unsuccessfully to other analytical methods.

The input and output density. The mean number of arrivals and departures per day of samples to and from the laboratory and different sections in the model are not significantly different from those in the actual laboratory; this follows from the Student's t values listed in Table 5. Kolmogorov—Smirnov tests, applied on the cumulative density functions of the input and output density of the simulated and actual laboratory, could not detect significant differences (Table 5). Subsequent G tests on the time series of the number of departures gave no significant difference for the residual variance and autocorrelation coefficient for the p.m.r., m.s. and ^{13}C-n.m.r. sections (Table 5).

TABLE 5

Mean, correlation at $\tau = 1$, and variance of output and input density of actual and simulated data, Student's t test on the means, $G(0,1)$ test on $\Phi(1)$ and residual variance, K—G test on cumulative density functions

Input	Actual data			Simulated data			t^a	$G(0,1)^b$	D^c
	γ	VAR(γ)	Φ_1	γ	VAR(γ)	Φ_1			
I.r.	2.8	13.2	0.13	2.6	5.3	0.14	0.74	*26.4	0.028
P.m.r.	7.7	20.3	0.18	7.2	12.7	0.42	1.0	*16.4	0.073
M.s.	2.1	4.2	0.08	1.8	1.9	0.12	1.5	*19.3	0.061
^{13}C-n.m.r.	2.5	4.2	0.02	2.5	3.0	0.0	0.15	3.0	0.054
Lab	11.8	44.1	0.14	10.8	24.2	0.51	1.4	*34.2	0.099
Output									
I.r.	2.8	7.5	0.0	2.6	20.8	0.07	0.66	*32.2	0.028
P.m.r.	7.7	50.0	0.0	7.2	67.9	0.0	0.68	2.78	0.066
M.s.	2.1	5.9	0.10	1.8	6.0	0.07	1.11	0.08	0.109
^{13}C-n.m.r.	2.5	11.1	0.0	2.5	11.3	0.0	0.08	0.007	0.096
Lab	11.8	77.6	0.0	10.8	146.9	0.0	1.04	*12.5	0.099

[a] $t_{0.005}(200) = 2.6$. [b] $1/2\ \chi^2_{0.01}(3) = 5.7$. [c] $D_{0.05} = 0.119$.

Conclusions

Queueing theory and digital simulation are valuable tools for understanding the mechanisms of queue formation in analytical laboratories. With queueing theory preliminary information can be obtained on the relative effects of various parameters and variables. The actual behavior of analytical laboratories can be digitally simulated.

The author thanks F. J. Spruit., F. W. van Deursen and all spectroscopists in the laboratory for structural analysis of Philips Duphar, B. V. Weesp, The Netherlands, for many helpful discussions and the collected data.

REFERENCES

1 D. L. Massart and L. Kaufman, Anal. Chem., 47 (1975) 1244A.
2 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
3 D. L. Massart, C. Janssens and L. Kaufman, Anal. Chem., 44 (1972) 2390.
4 D. L. Massart, L. Kaufman and R. Smits, Z. Anal. Chem., 264 (1973) 273.
5 B. Schmidt, GIT Fachz. Lab., 20 (1976) 1167.
6 B. Schmidt, Fresenius Z. Anal. Chem., 287 (1977) 157.
7 I. K. Vaananen, S. Kivirikko, J. Koskenniemi, J. Koskimies and A. Relander, Meth. Inform. Med., 13 (1974) 158.
8 V. Adeberg and K. Doerffel, Euroanalysis II, Budapest (1975).
9 J. R. Jackson, Oprns. Res., 5 (1957) 518.
10 Y. Levy and U. Yechiali, Manag. Sci., 22 (1975) 202.
11 M. J. Fisher, Manag. Sci., 24 (1977) 338.
12 L. Kleinrock, Queueing Systems, Vol. II, Wiley, New York, 1976, pp. 119—126.

13 L. Kleinrock and P. Finkelstein, Oprns. Res., (1976) 104.
14 R. W. Conway, W. L. Maxwell and L. W. Miller, Theory of Scheduling, Addison-Wesley, Reading, Mass., 1967.
15 D. A. Hsu and J. S. Hunter, Manag. Sci., 24 (1977) 181.
16 G. E. P. Box and G. M. Jenkins, Time Series Analysis Forecasting and Control, Holden-Day, San Francisco, 1970.
17 G. E. P. Box and G. C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, Mass., 1973.
18 J. A. G. M. Kerbosch and R. W. Sierenberg, Discrete simulatie m.b.v., ALGOL, FORTRAN, PL/1, Samson, Brussels, 1973.
19 A. J. Lemoine, Manag. Sci., 24 (1977) 464.

Analytica Chimica Acta, 112 (1979) 277—286
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

EINSATZ VON RECHENANLAGEN IN DER EMISSIONSSPEKTROMETRIE DEFINITION DER PROBLEME UND DIE SCHWÄRZUNGSTRANSFORMATION*

MIKULÁŠ MATHERNY

Lehrstuhl für Chemie der Hüttenmännischen Fakultät der Technischen Hochschule, Svermova 9, CS-043 85 Košice, (Czechoslovakia)

(Eingegangen den 13 Februar 1979)

ZUSAMMENFASSUNG

Das Problem der Bestimmung der von systematischen Fehler freien Parametern der Schwärzungstransformation sowie die eigentliche Schwärzungstransformation beinhaltet eine Reihe von Operationen. Bei der Ermittlung der Transformationsparameter ist es die modifizierte Churchill'sche Vorkurve, die die rasche Berechnung des γ-Wertes ermöglicht, aber die Festlegung der wahren Transformationskonstante k der l-Transformation muss durch iterative Annäherung verbessert werden. Gerade diese Operation erlaubt aber gleichzeitig auch eine genaue Ermittlung des κ-Wertes der allgemeinen P-Transformation. Bei beiden Ermittlungsarten der Transformationskonstanten ist das Bestimmungsverfahren mit einer Zuverlässigkeitsprüfung verbunden, die den vom systematischen Fehler freien Charakter der Berechnungen sichert.

SUMMARY

The Use of Computers in Emission Spectrometry. Definition of the Problem and the Blackening Transformation.

The problems of determining the parameters of the blackening transformation free of systematic errors, as well as the blackening transformation itself, are discussed. With regard to the transformation parameters, the modified Churchill's preliminary curve enables fast calculation of the γ-value (slope), although the accurate transformation constant (k) of the l-transformation has to be determined by an iterative approach. It is also possible to obtain highly precise κ-values for the general P-transformation by iteration. In both methods of obtaining the transformation constants, the procedure involves confidence tests which ensure that the calculation is free from systematic errors.

Der Einsatz von Rechenanlagen in der atomaren Emissionsspektrometrie unter Anwendung der photographischen Detektion ist hauptsächlich bei der Ausarbeitung der analytischen Methoden mit einem statistischen Optimierungsverfahren verbunden [1—3]. Die Probleme der Optimierung werden zweckmässig in folgender Reihenfolge diskutiert.

(A) Ermittlung der Schwärzungswerte, Festlegung der Transformations-

*Vorgetragen am Colloquium Spectroscopicum Internationale, Štrbské Pleso, September 1977.

parameter und eigentliche Transformation der Schwärzungswerte in die *Y*-Werte und eventuell auch in relative Intensitätswerte *I*.

(B) Komplexe analytische Eichung und Bewertung dieser Prozedur, die eigentlich aus drei zusammenhängenden Teilen besteht: die eigentliche Aufstellung der linearen analytischen Eichgeraden sowie ihre Optimierung, die Ermittlung der Nachweis- und Garantiegrenzen, und schliesslich die Berechnung der "totalen" relativen Präzision der Konzentrationsbestimmung.

(C) Optimierungsverfahren der atomaren Emissionsspektrometrie, die mit der Aufdeckung und Beseitigung der Matrixeffekte beginnen, mit der Überwachung der Verdampfungsprozesse fortgesetzt werden und letztlich mit der Auswahl der analytischen Linienpaare enden. Mit diesen Aufgaben hängen auch die Probleme der Temperatur-, Elektronendruck- und Teilchenkonzentrationsbestimmung eng zusammen.

(D) Anwendung der Rechenanlagen für die eigentliche Konzentrationsbestimmung sowie eine programmierte Ausgabe dieser Daten in Verbindung mit der Wertung der Endergebnisse grösserer Matrices von Konzentrationswerten.

Die definierten Probleme berücksichtigen die Eigentümlichkeiten der Pulver- und Lösungsanalysen, die meistens mittels klassischer Anregungsquellen angeregt werden. Bei diesen Quellen muss man mit einer bedeutenden Fluktuation der analytischen Signale rechnen, die immer höher sind, als die Signale der Atomabsorption.

ANWENDUNG DER RECHENANLAGEN BEI DER SCHWÄRZUNGSTRANSFORMATION

Die Schwärzungstransformation verlangt die Bestimmung der korrekten Parameterwerte der angewendeten Transformationsoperation. Zur Zeit haben sich zwei Arten von Transformationen in der emissionsspektrometrischen Praxis eingebürgert: es ist die von Török und Zimmer [4] vorgeschlagene *l*-Transformation und die allgemeine *P*-Transformation nach Kaiser [5]. Beide Transformationen besitzen einen empirischen Charakter und verlangen eine Eichung, die die Werte der Transformationsparameter liefert, und zwar einerseits den γ-Wert, der die Richtungstangente des geradlinigen Teils der Eichkurve der photographischen Emulsion (Schwärzungskurve) darstellt, und anderseits die eigentlichen Transformationskonstanten. Für die *l*-Transformation ist das der *k*-Wert und für die *P*-Transformation der κ-Wert. Die Autoren der genannten Transformationsoperationen haben auch die entsprechenden Formeln für die Berechnung der γ-, *k*- und κ-Werte definiert, doch macht die Streuung der direkten Berechnung dieser Werte [3, 4, 6] eine iterative Verbesserung sowie eine nachfolgende Zuverlässigkeitsprüfung [7] nötig. Hier soll angemerkt werden, dass auch in den letztpublizierten Arbeiten, die die Schwärzungstransformationsprobleme lösen ebenso wie in früheren, die Kontrolle der Zuverlässigkeit der ermittelten Werte fehlt [8—10].

Theoretische Überlegungen

Die Bestimmung der Transformationsparameter erfordert eine Anfangs-

matrix der $\{S_{b,i}, S_{a,i}\}$-Schwärzungswertepaare, wobei $S_{b,i} > S_{a,i}$ $(i = 1, \ldots N)$ und $N \in \langle 20, 100 \rangle$ ist. (Die angewendeten Symbole sind in Tabelle 1 erklärt.) Diese Wertepaare erhält man aus Spektren, die mit einem Zweistufenfilter mit einer Filterkonstante ΔY_m aufgenommen werden. Der funktionelle Zusammenhang zwischen den S_b- und S_a-Werten durch

$$S_{a,i} = B_i S_{b,i} - r_i \tag{1}$$

ist gegeben. Im geradlinigen Teil der von Plško [11] modifizierten Churchill'schen [12] Vorkurve (Abb. 1) weist der Wert r, der den Schnittpunkt mit der S_b-Achse der Beziehung (1) darstellt, einen konstanten Wert auf und B muss logischerweise gleich Eins sein. Im gekrümmten Teil dagegen gilt $r_i = f(S_b)$ sowie $B_i = f(S_b)$. Weiter gelten aus logischen Überlegungen [11] auch die weiteren zwei Gleichungen

$$\gamma = r/\Delta Y_m \tag{2}$$

$$k \equiv s_l = S_L/\gamma \tag{3}$$

Die rechnerische Aufgabe liegt also in einer möglichst vom systematischen Fehler freien Ermittlung der γ- und S_L-Werte sowie in einer nachfolgenden Zuverlässigkeitsprüfung. Diese Prozedur [13] basiert auf der Idee [14], dass die ΔY-Werte, die aus den S_b- und S_a-Werten stammen, die für die Eichung der photographischen Emulsion verwendet wurden, für die Gesamtzahl der Messwerte mit dem Werte der Filterkonstante ΔY_m des Zweistufenfilters im Rahmen der Messgenauigkeit übereinstimmen müssen:

$$(1/N) \sum_{1}^{N} (\Delta Y_i - \Delta Y_m) = 0 \tag{4}$$

TABELLE 1

Erläuterung der angewendeten Symbole

I	relativer Intensitätswert einer Spektralinie
ΔY	Differenz von zwei Y-Werten; $Y = \log I$
S	gemessener Schwärzungswert
$S_{L,T}$	berechneter S_L-Wert
s	sog. reduzierter Schwärzungswert
γ	Richtungstangente des geradlinigen Teils einer Eichkurve der photographischen Emulsion
γ_P	Richtungstangente der P-transformierten Eichkurve der photographischen Emulsion
k	Transformationskonstante der l-Transformation
κ	Transformationskonstante der P-Transformation
ΔY_m	Wert der Filterkonstante eines Zweistufenfilters
D	$= S - W$; $D_s = s - W_s$
W	sog. Gauss'scher Subtraktionslogarithmus
MIN	Minimumkriterium
N, M, K, Q	Endzahl von Werten
i, j, q	Reihenfolge von Messwerten

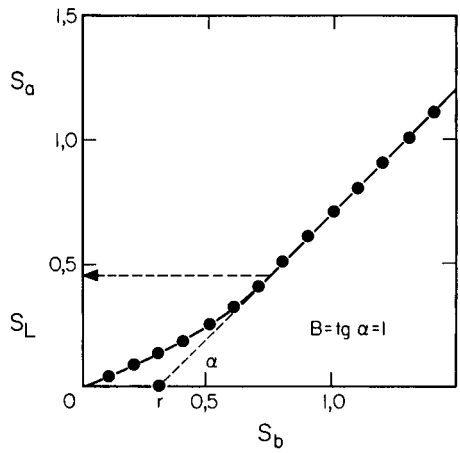


Abb. 1. Modifizierte Churchill'sche Vorkurve.

Die Berechnung der Transformationskonstante κ aufgrund der Summationsformel

$$\kappa = \left[1/K \sum_1^K eS_q - 1/Q \sum_1^Q eS_j\right] \Big/ \left[1/K \sum_1^K eD_q - 1/Q \sum_1^Q eD_j\right] \tag{5}$$

führt manchmal gerade durch fehlerhafte Auswahl der S_q- und S_j-Schwärzungswerte zu falschen Resultaten. Es sind die jene Fälle in denen die $S_{a,q}$-Werte für die Bildung der eS_q-Differenzen (siehe unten) nicht ausschliesslich aus dem linearen Verlauf der Eichkurve der photographischen Emulsion stammen, oder wo die $S_{b,j}$-Werte für die Bildung der eS_j-Differenzen nicht ausschliesslich aus dem gekrümmten (unterexponierten) Teil ausgewählt wurden.

$$eS_q = S_{b,q} - S_{a,q}; \; eS_j = S_{b,j} - S_{a,j}$$

$$eD_q = D_{b,q} - S_{a,q}; \; eD_j = D_{b,j} - D_{a,j}$$

Deshalb muss der κ-Wert entsprechend Gl. (5) nur aus der reduzierten Matrix der $\{S_{b,i}, S_{a,i}\}$-Schwärzungswertepaare berechnet werden. Für die Differenzwerte eS_q dürfen nur solche Wertepaare verwendet werden, die der Ungleichung $S_{a,i} \geqslant S_{L,T} + 0.05$ genügen, und für die Differenzwerte eS_j nur solche Wertepaare, die der Ungleichung $S_{b,i} \leqslant S_{L,T} - 0.05$ entsprechen. Den $S_{L,T}$-Wert erhält man durch Modifikation der Gl. (3): $S_{L,T} = k\gamma$.

Berechnung der Transformationsparameter der l-*Transformation*

Der eigentliche Rechengang (Abb. 2) beginnt im ersten Block mit der Abstufung der Glieder der Schwärzungswertepaare in sinkender Reihenfolge der $S_{b,i}$-Werte. Für die genannten Wertepaare werden, unter Anwendung der Methode der kleinsten Quadrate nach dem Minimalisierungsprinzip

$$1/N \sum_1^N [B_i \cdot S_{b,i} - r_i - S_{a,i}]^2 = \epsilon = \text{MIN} \tag{6}$$

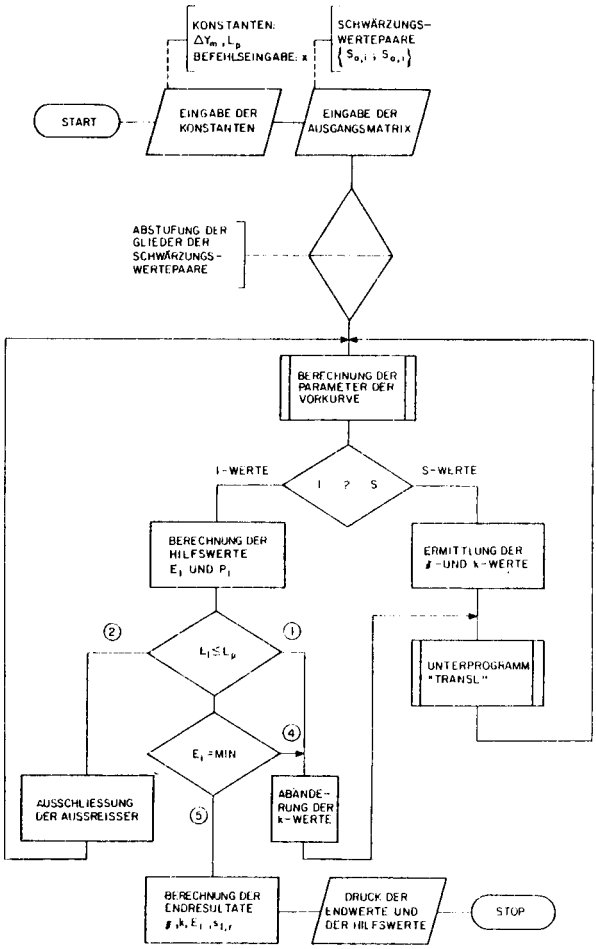


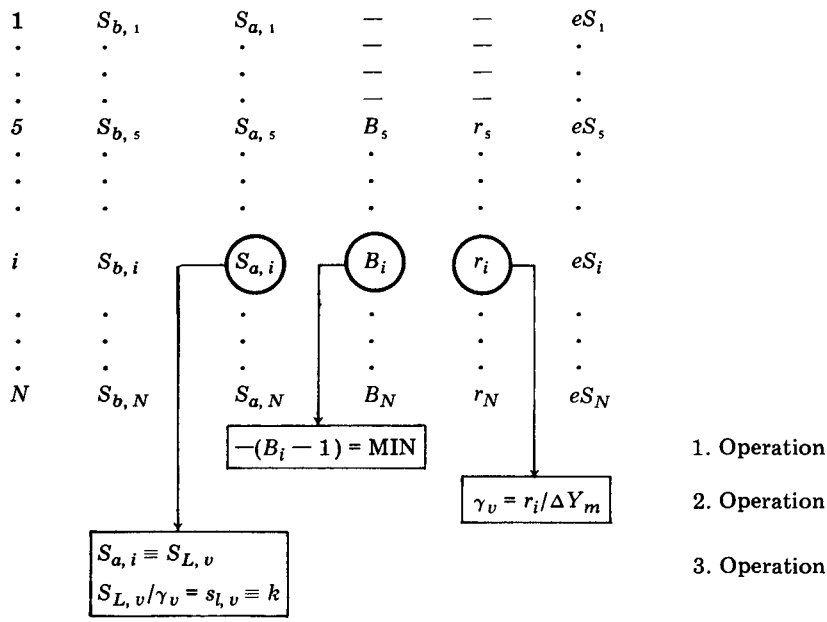
Abb. 2. Flussdiagram der Ermittlung der Transformationsparameter der *l*-Transformation; 1. bei $E_l > L_p$, 2. bei $E_l \leqslant L_p$, 3. nach der Ausschliesungsprozedur, 4. bei $E_{l,\,i+1} < E_{l,\,i}$, 5. bei $E_{l,\,i+1} \geqslant E_{l,\,i}$.

die B_i-, r_i- sowie die eS_i-Werte ($eS_i = S_{b,\,i} - S_{a,\,i}$) berechnet. Die ersten B- und r-Werte werden aber, zur Beseitigung eventueller Störfaktoren gleich von den ersten fünf Schwärzungswertepaaren berechnet ($i = 5$), und die weiteren durch die sukzessive Hebung der Zahl der Wertepaare bis $i = N$ (Tabelle 2). Dieser $S_{a,\,i}$-Wert, bei dem für die B_i-Werte die minimale negative Abweichung ($B_i < 1$) von Eins erreicht wird, wird als vorläufiger $S_{L,\,v}$-Wert betrachtet. Der zugehörige r_i-Wert wird für die erste Berechnung des vorläufigen γ_v-Wertes verwendet Gl. (2), und der gesuchte vorläufige k_v-Wert ergibt sich aus Gl. (3) (Tabelle 2, Operationen 1, 2 und 3).

Anhand der ermittelten vorläufigen Transformationsparameterwerte werden

TABELLE 2

Feld der Schwärzungswertepaare und zugehörige Hilfswerte der modifizierten Vorkurve



im weiteren alle Glieder der Anfangsmatrix, nach den Regeln der l-Transformation [4] in die l-Werte umgerechnet:

$$l = s - d(k - s) \tag{7}$$

wo $s = S/\gamma$ und $D_s = s - W_s$. Selbstverständlich werden nur diese s_i-Werte transformiert, die die Ungleichheit $s_i < s_l$ erfüllen, denn für Werte $s_i > s_l$ gilt $s_i \equiv l_i$. Gleichzeitig werden mit der schon erörterten Minimalisationsprozedur (Gl. 6) unter Austausch der S-Werte für die transformierten l-Werte die neuen B_i-, r_i- sowie el_i-Werte berechnet (Tabelle 3). Aus den el_i-Werten bildet man im weiteren den Mittelwert sowie die entsprechende Standardabweichung:

$$\overline{el} = (1/N)\sum_{1}^{N} el_i \tag{8}$$

$$s_{el} = \left\{\left[\sum_{1}^{N} (el_i - \overline{el})^2\right] \Big/ [N-1]\right\}^{\frac{1}{2}}$$

Die Verlässigkeit der Bestimmung der Transformationsparameterwerte wurde mittels der Hilfswerte E_1 und P_l geprüft:

$$E_l = 1/N \sum_{1}^{N} (el_i - \Delta Y_m) \tag{9a}$$

$$P_l = 1/N \sum_{1}^{N} (el_i/\Delta Y_m) \tag{9b}$$

TABELLE 3

Feld der transformierten l-Werte und zugehörigen Hilfswerte der modifizierten Vorkurve

1	$l_{b,1}$	$l_{a,1}$	B_1	r_1	el_1	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
M	$l_{b,M}$	$l_{a,M}$	B_M	r_M	el_M	
				$\gamma = r_M/\Delta Y_m$		1. Operation
				$E_l = \frac{1}{M}\sum_{1}^{M}(el_i - \Delta Y_m)$		2. Operation

Diese Hilfswerte werden aber ausschliesslich von solchen $\{l_{b,i}, l_{a,i}\}$-Wertepaaren berechnet, für welche die Ungleichheit $l_{b,i} < k$ gilt [15]. Falls E_l gleich Null und P_l gleich Eins sind, besitzen die als vorläufig behandelten Parameterwerte wahre Werte, und dadurch erlangt die Berechnung ihre Verlässigkeit.

Falls aber E_l von Null abweichende Werte aufweist, was beim ersten Schritt der häufigste Fall ist, muss eine iterative Annäherung zum wahren k-Wert eingeleitet werden. Entscheidend für die weiteren Operationen ist, ob $E_l \leqslant 0.01$ ist. In diesem Fall wird der k_v-Wert nur um den Schritt 0.01 verändert. Falls $P_l < 1$ ist, wird der neue k-Wert um diesen Schrittwert gesenkt; falls aber $P_l > 1$ ist, wird er um diesen Schrittwert erhöht. Für $E_l > 0.01$ ist es günstiger, einen Schrittwert von 0.02 zu benützen und so die Zahl der Iterationen zu vermindern. Nach der Veränderung des k-Wertes werden neue $l_{b,i}$- und $l_{a,i}$-Werte sowie die el_i-Werte gebildet und ihre Zuverlässigkeit wird von neuem geprüft. Diese Iterationsprozedur läuft bis $E_l \leqslant L_p$. Der L_p-Wert muss jeweils beim Start des Rechenganges als Eingangsgrösse eingegeben werden, und stellt eigentlich das Zuverlässigkeitskriterium dar. Meistens arbeitet man mit Werten zwischen 0.01 und 0.005.

Nach diesem Komplex von Operationen müssen aus der transformierten $\{l_{b,i}, l_{a,i}\}$-Matrix noch diejenigen Wertepaare ausgeschlossen werden, deren el_i-Wert nicht in die Spannweite fällt. Der Multiplikationsfaktor x muss

$$(\overline{el} \pm x s_{el})$$

ebenso wie die anderen bereits erwähnten Werte im Voraus angegeben werden. Für den Multiplikationsfaktor x hat sich der Nominalwert 1.55 bewährt, da bei der Anwendung dieses Kriteriums ca. 10% der el_i-Werte, die sogenannten stark störende Werte, ausgeschlossen werden.

Im zweiten Block, mit der reduzierten Zahl $M(M < N)$ der $\{S_{b,i}, S_{a,i}\}$-Wertepaare (wo $i = 1, \ldots, M$) und mit den oben definierten komplexen

Prozeduren werden der γ-Wert sowie der vorläufige k_v-Wert von neuem berechnet. Dieser wird in der schon definierten Iterationsprozedur sukzessiv verbessert. Das Ende der Iterationen wird aber nicht durch die Ungleichheit $E_l \leqslant L_p$ gegeben, sondern diese laufen bis der E_1-Hilfswert durch die sukzessive Änderung der k-Werte sinkt. Die Zuverlässigkeit der Transformation mit dem endgültigen k- und γ-Wert ist nun durch die erreichten r_M-und E_l-Werte durch den Mittelwert el und durch die Standardabweichung s_{el} gegeben. Diese Werte muss man aber in diesem Fall aus allen Gleidern der reduzierten $\{l_{b,i}, l_{a,i}\}$-Matrix berechnen (Tabelle 3, Operationen 1, 2). Die relative Präzision der Prozedur ist folgendermassen definiert: $s_{l,r} = (s_{el}/\overline{el})\ 100$.

Berechnung der κ-Transformationskonstanten

Die Berechnung der κ-Konstante der allgemeinen P-Transformation mittels Summationsformel (Gl. 5) erfordert erstens die Berechnung der sogenannten theoretischen $S_{L,T}$-Werte (= $k\gamma$) und eine Aufteilung der ursprünglichen, schon reduzierten Anfangsmatrix in zwei getrennte Matrixen $\{S_{b,q}, S_{a,q}\}$ und $\{S_{b,j}, S_{a,j}\}$, die die Trennungskriterien $S_{a,i} \geqslant S_{L,T} + 0.5; S_{b,i} \leqslant S_{L,T} - 0.5$ in Betracht ziehen. Im weiteren wird aufgrund der Gl. (5) der vorläufige κ_v-Wert berechnet, und mit diesem Wert werden alle Schwärzungswerte entsprechend Gl. (10) in die P-Werte umgerechnet:

$$P = S - \kappa_v \cdot D \tag{10}$$

Die Matrix $\{P_{b,i}, P_{a,i}\}$ wird mittels Gl. (6) bearbeitet, aber anstelle der S-Werte werden bereits die transformierten P-Werte verwendet. Die Zuverlässigkeit des κ_v-Wertes wird anhand der berechneten eP_i Differenzwerte, beziehungsweise deren arithmetischen Mittelwert $\overline{eP}$ beurteilt. Falls nach einem Iterationskreis der Zustand

$$E_P = (\overline{eP} - \gamma \Delta Y_m) = \text{MIN} \tag{11}$$

erreicht wird, endet die Iterationsprozedur. Jener r_i-Wert, der aus derjenigen Matrix stammt, beider das Minimumkriterium (Gl. 11) erreicht wurde, wird für die Ermittlung des γ_P-Wertes nach Gl. (2) herangezogen, und der entsprechende κ-Wert wird als endgültige Transformationskonstante behandelt.

SCHWÄRZUNGSTRANSFORMATION

Falls man über einen zuverlässigen γ- oder γ_P-Wert verfügt, und entweder die k- oder aber die κ-Transformationskonstanten vorhanden sind, wird im Weiteren mit diesen Werten und mit einer der beiden diskutierten Operationen gearbeitet. Die eigentliche Schwärzungstransformation [4, 5] ist vom Gesichtspunkt der Programmierung eine triviale Aufgabe. Es handelt sich eigentlich nur um die programmierte Lösung der Grundgleichungen (7) und (10). Dementsprechend wurden die zwei Transformationsoperationen als fixe Unterprogramme zusammengestellt. Für die weiteren Rechenprogramme dieser Serie von Arbeiten wurde nur das Unterprogram TRANSL, welches für die l-Transformation bestimmt ist und in FORTRAN IV geschrieben ist, angewendet.

Für die allgemeine P-Transformation muss man aber die transformierten P-Werte mit dem bestimmten γ_P-Wert dividieren um so die gesuchten Y-Werte zu erhalten. Unter Berücksichtigung dieser Operation gilt weiter [16] dass $\Delta l \equiv \Delta Y$, obwohl sich die transformierten l- und Y-Werte unterscheiden. Das entsprechende Unterprogram TRANSP ist wiederum in FORTRAN IV geschrieben (Abb. 3).

Schlussbemerkung

Die im Vorhergehenden beschriebenen Rechenverfahren sind in den im folgenden genannten 5 FORTRAN IV Programmen codiert: Unterprogramm TRANSL für die l-Transformation der Schwärzungswerte; Unterprogramm TRANSP für die allgemeine P-Transformation der Schwärzungswerte; Pro-

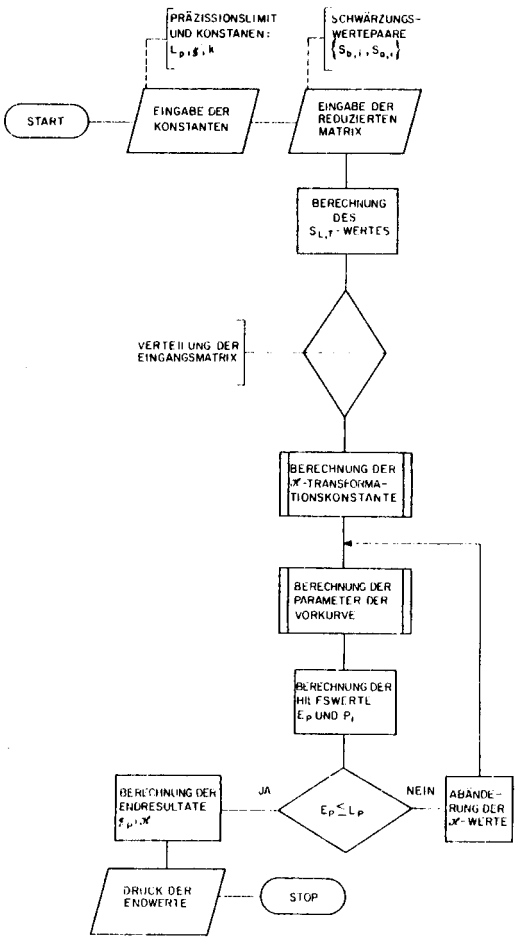


Abb. 3. Flussdiagram der Ermittlung der Transformationskonstante κ der allgemeinen P-Transformation.

gramm DEPALT G/K-M-78 für die Ermittlung der Werte der Transformationsparameter der *l*-Transformation; Programm DEPAPT GP/KAPA-M-78 für die Ermittlung der Werte der Transformationsparameter der allgemeinen *P*-Transformation; Programm DEPALPT G/K-GP/KAPA-M-79 für die gleichzeitige Ermittlung der Werte der Transformationsparameter der *l*-Transformation sowie der allgemeinen *P*-Transformation.

Diese Programme stehen Interessenten zur Verfügung.

Der Verfasser dankt Frau H. Horovčáková vom Rechenzentrum der TH-Košice für die Hilfeleistung bei der Abfassung der Programme, und Herrn Dr.-Ing. K. Flórián vom Lehrstuhl für Chemie der TH-Košice für Diskussionen zum bearbeiteten Problemkreis.

LITERATUR

1 M. Matherny, Kém. Köz., 48 (1977) 363.
2 M. Matherny, Mitteilungsheft Chem. Gesell. DDR, im Druck.
3 M. Matherny, Wiss. Z. K. Marx-Univ., Math.-Naturwiss. Reihe, im Druck.
4 T. Török und K. Zimmer, Quantitative Evaluation of Spectrograms by Means of *l*-Transformation, Akadémiai Kiadó, Budapest and Heyden, London, 1972.
5 H. Kaiser, Spectrochim. Acta, 3 (1947) 159.
6 K. Zimmer und T. Török, Acta Chim. (Budapest), 29 (1961) 273.
7 K. Flórián und M. Matherny, Spectrochim. Acta, 33B (1978) 429.
8 M. Malínek, Can. Appl. Spectrosc., 20 (1975) 68.
9 J. A. Holcombe, D. W. Brinkman und R. D. Sacks, Anal. Chem., 47 (1975) 441.
10 D. R. Blevins und W. R. O'Neill, Appl. Spectrosc., 30 (1976) 190.
11 E. Plško, Chem. Zvesti, 23 (1969) 150.
12 J. R. Churchill, Ind. Eng. Chem., Anal. Ed., 16 (1944) 653.
13 M. Matherny, Kém. Köz., 48 (1977) 365.
14 K. Laqua, W. D. Hagenah und U. Haisch, Proc. XVth. Coll. Spectrosc. Internat., Madrid, 1969, Vol. IV, p. 380.
15 K. Flórián, Gy. Heltai und K. Zimmer, Acta Chim. (Budapest), im Druck.
16 T. Török und K. Zimmer, Annales Univ. Sci., Budapest, Sect. Chim., 8 (1966) 11.

Analytica Chimica Acta, 112 (1979) 287—295
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHARACTERIZATION OF THE FLOW DYNAMICS OF AN ENZYME REACTION SYSTEM

CHARLES L. POMERNACKI, THOMAS A. BRUBAKER**, HAL R. BRAND and JACK W. FRAZER*

Lawrence Livermore Laboratory, University of California, Livermore, CA 94550 (U.S.A.)

(Received 16th March 1979)

SUMMARY

The characterization of the flow dynamics of an automated enzyme reaction system is discussed. Parametric and non-parametric characterizations are constructed and then tested against actual experiments. The analytical model provides a more compact representation than the numerical model. A second-order lag-plus-dead-time model gives a good fit to the experimental data.

Complete characterization of chemical processes is normally very difficult and time-consuming. With properly designed apparatus and pilot plants that incorporate appropriate chemical instrumentation, vastly improved chemical information can be obtained. When such systems also include on-line computers and graphic display equipment, complete characterization becomes faster and easier.

Considerable research is required to develop on-line instrumentation and the supporting signal processing, modeling, and display techniques. An experimental test apparatus (Fig. 1) has been built to study the automated characterization of complex non-linear chemical systems, and is currently being used to study enzyme-catalyzed reactions. The general features of the apparatus, graphic techniques, and preliminary chemical data have been described [1, 2].

The flow dynamics of the system must be characterized for three reasons. First, it is necessary to know when the system has reached a steady-state in order to determine when to begin taking the data that will be used to calculate the static values that form response surfaces. Response surfaces are used to obtain a rapid picture of the equilibrium steady-state kinetics. Secondly, for very rapid characterization of a chemical reaction it may be desirable to make the measurements before the system reaches the steady state. Thirdly, dynamic characterization is needed in the design procedure to develop an effective process controller.

**Present address: Department of Electrical Engineering, Colorado State University, Fort Collins, Colorado 80523.

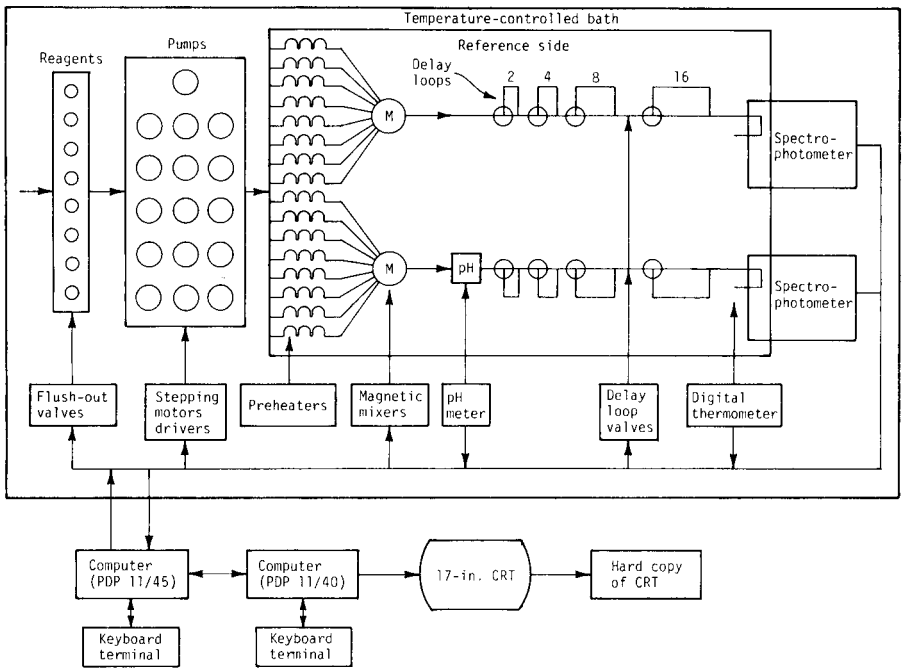


Fig. 1. Block diagram of the computerized apparatus showing the flow system and major components. Each reagent is aspirated from its reservoir, pumped through a preheater, and mixed with other reagents to initiate the reaction. The mixture is then pumped through the reaction delay tubes to the spectrophotometers.

The term characterization implies that a descriptive and representational process is involved in understanding the system. Characterization involves defining the system elements, developing an experimentation plan, acquiring the data, modeling the input and output transducers, and modeling the process. Modeling implies development of a copy or imitation of the system that predicts the behavior of a system. Models are useful in analyzing, understanding, and controlling the process. There are two classes of models that can be used to characterize a deterministic, linear, dynamic system. The first is a parametric or analytical model such as a transfer function or state-space model. The second is a non-parametric or numerical data model. Both models are discussed in this paper.

In characterizing the flow dynamics of the enzyme system, the topic of concern was only how long it takes fluids to move through the apparatus, not the reaction kinetics. There are several methods that can be used to characterize dynamic systems, but all methods consist of two parts. First, the nature of the excitation signal must be established. This signal can be a periodic function, a noise source, an approximate impulse, or a step function. A step

function was employed here because it appears to have advantages over both the periodic and noise signals in systems exhibiting large time constants. In addition, a step function avoids the unrealistic need, as in the case of the impulse signal, to drive the system inputs to large amplitudes. Secondly, a computational procedure must be chosen. Computational procedures such as least squares and deconvolution are available for operation on numerical step-response data. The least-squares approach is used to determine the parameter values in a transfer function model by using the corresponding time-domain step-response expression. The deconvolution approach is used to develop a numerical model from the step-response data. Numerical deconvolution techniques such as computing the Fourier transform, solving the convolution integral equation, and differentiating the step response are all viable [3]. Differentiation of the step-response signal was selected for this work because of the existence of a robust, computationally efficient numerical differentiation procedure (see below).

CHARACTERIZATION BY DECONVOLUTION TECHNIQUES

The process of characterizing a dynamic system is known as system identification [3]. One way of stating the identification problem is [4]: given an input $s(t)$ (in the present case, a step function) and an output $r(t)$, find the convolution function (impulse response or Green's function) $f(t)$. Once $f(t)$ is known, the output, $r(t)$, can be found for any input, $s(t)$. The task is to find the convolution function from the input—output data.

In this work, the convolution function takes two forms. The first is the numerical-non-parametric form of the convolution function, which can be constructed directly from the output data without using analytical expressions. The second is the analytical-parametric form, which requires selection of parameters in a proposed analytical form of the convolution function. Thus the system response can be described analytically and the results of the analytical expression compared directly to the output data set.

Let us briefly review the convolution operation. The convolution operation is defined as

$$r(t) = \int_0^t f(\lambda)\, s(t-\lambda)\mathrm{d}\lambda \tag{1}$$

To obtain $f(t)$, a deconvolution procedure is utilized with a unit step input. If it is assumed that $s(t) = 1$ for $t \geqslant 0$, and $s(t) = 0$ for $t < 0$, then substituting into eqn. (1) yields the expression

$$r(t) = \int_0^t f(\lambda)\mathrm{d}\lambda \tag{2}$$

Differentiating both sides of this equation with respect to time and applying a basic theorem of calculus yields

$$\frac{dr(t)}{dt} = \frac{d}{dt}\int_0^t f(\lambda)d\lambda = f(t) \tag{3}$$

where $f(t)$ is continuous over the interval $(0, t)$. Thus, the derivative of the step response is the desired convolution function.

Given sampled data taken from the linear system response, the convolution function can be obtained numerically by using a robust (noise-insensitive) numerical differentiation algorithm. A well-known algorithm for accomplishing this has been described by Savitzky and Golay [5]. The 19-point algorithm chosen is based on a 6th-degree polynomial that yields an estimate for the derivative. The convolution function, at discrete times, is given by

$$r[kT] = \sum_{i=0}^{9} C_i\,[s(kT + iT) - s(kT - iT)]/N{\cdot}T \tag{4}$$

where $C_0 = 0$, $C_1 = 332684$, $C_2 = 583549$, $C_3 = 686099$, $C_4 = 604484$, $C_5 = 348823$, $C_6 = -9473$, $C_7 = -322378$, $C_8 = -349928$, $C_9 = 255102$, $N = 9806280$, $T = 1$ s, i = summation index, and k = sampling index.

When eqn. (4) is used, the algorithm must be started far enough into the data set so that the 19-point window is filled. Otherwise undesirable transient effects can occur.

A non-linear least-squares technique was used to obtain the parameters of a proposed analytical convolution function given the sampled step-response data. The proposed analytical form of the convolution function was then utilized, and the right-hand side of eqn. (2) was analytically integrated. The least-squares technique was then used to choose the parameters of the integrated convolution function, i.e. the step response, that gave the best fit to the step-response data.

EXPERIMENTAL RESULTS

This deconvolution procedure was tested on an alkaline phosphatase-catalyzed reaction. The experimental setup is outlined in Fig. 2, which is a simplified schematic diagram of the experimental apparatus (Fig. 1) showing the material and information flow paths. In a typical experiment, a step input of enzyme is injected by controlling the appropriate pump. The quantity of product (the output) is then measured. The response represents the entire flow dynamics in the material flow path including the effect of the spectrophotometer. The pump response for injecting enzyme is much faster than the flow dynamics and can be neglected. The following experimental conditions were used for the alkaline phosphatase-catalyzed reaction: pH 10.0; substrate concentration 3 mmol l^{-1}; temperature 30.0°C; data sampling rate 1.0 s; alkaline phosphatase concentrations 3 and 12 IU l^{-1}.

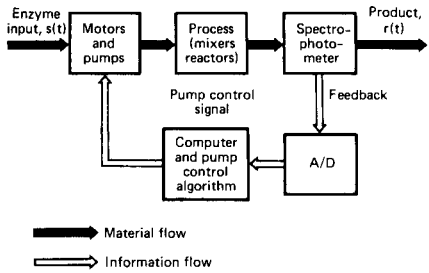


Fig. 2. Block diagram of the enzyme control system showing the material and information flow paths.

Non-parametric model

The system was first stimulated with a step in enzyme concentration from 3 to 12 IU l^{-1} at time $t = 0$. A step in enzyme concentration was chosen because of the relationship between enzyme concentration and product concentration. The enzyme step input, $s(t)$, and the resulting step response in product concentration, $r(t)$, are shown in Fig. 3. The data are displayed as a continuous curve by linear interpolation between the data points. The initial concentration of product is approximately 4 μmol l^{-1}, the initial rise starts after a 60-s delay, the leading edge is smoothed out at about 180 s, and the final steady-state concentration level is about 17 μmol l^{-1}.

The numerical deconvolution technique is then applied to the response data and the convolution function is generated as shown in Fig. 3. Since all the dynamic behavior of the system is contained in the convolution function, the quality of this method can be tested by applying a numerical enzyme input of 3—12 IU l^{-1} to the convolution function and then observing the reconstructed step response output $r(t)$. Figure 4 shows an overlay of the original data on the reconstructed data. The curves match better than the line resolution of the graph, which verifies the accuracy of the convolution function.

The above procedure was repeated for a downward step in enzyme concentration from 12 to 3 IU l^{-1}. The original data for the step-response output product concentration and the corresponding convolution function are shown in Fig. 5. The two convolution functions (Figs. 3 and 5) are almost identical considering the sensitivity to noise of the numerical differentiation scheme.

To validate the method, it is necessary to apply it to other situations. A 40-s input pulse in enzyme concentration of 3 to 12 to 3 IU l^{-1} was applied to the convolution function with the result shown in Fig. 6. The convolution function previously determined was used to compute the theoretical pulse response. The predicted response and the actual response match well (Fig. 6). A summary of the methodology is shown in Fig. 7.

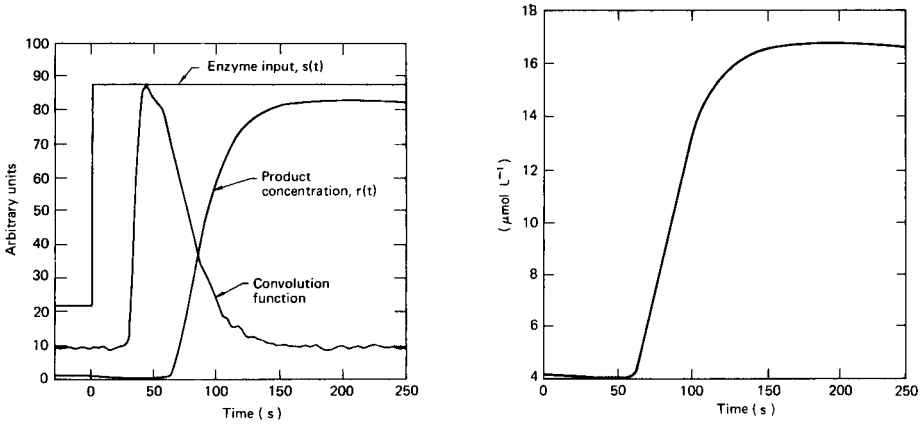


Fig. 3. The step input change in enzyme concentration, the resultant output change in product concentration caused by the step input change in enzyme concentration, and the convolution function.

Fig. 4. Comparison of the calculated and experimental response to a step change by using a second-order lag-plus-dead-time model, showing a near-perfect fit.

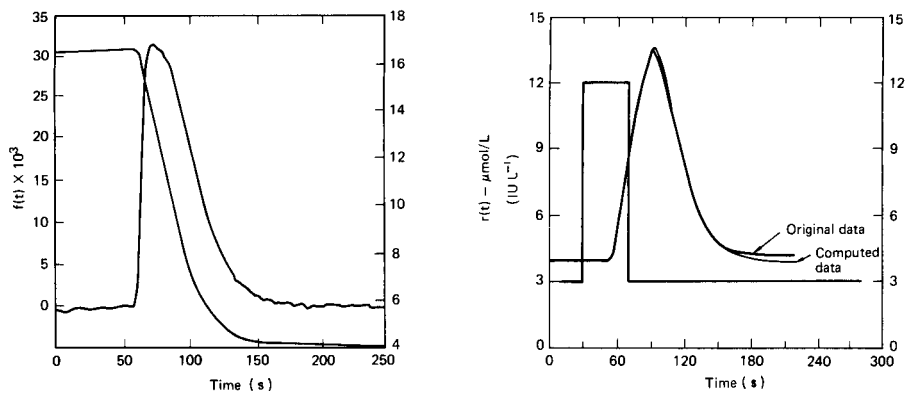


Fig. 5. Graphic representation of the convolution function and the resultant output change in product concentration caused by a downward step input change in enzyme concentratio

Fig. 6. Pulse (40 s) input change in enzyme concentration and a comparison of the calculated and actual system response to the enzyme pulse.

Parametric model

The analytical or parametric approach is actually only pseudo-analytical in that the mathematical forms are not actually derived from first principles, but instead tend to describe the dominant characteristics of a large class of such processes [6].

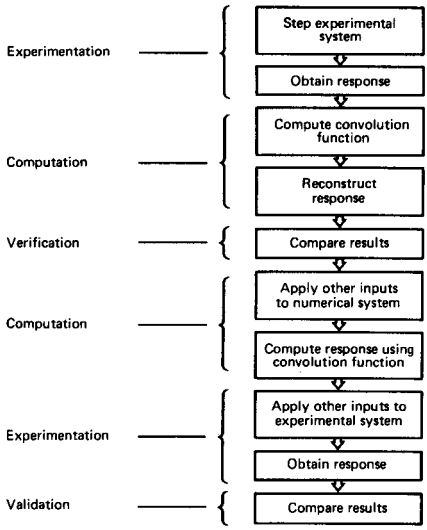


Fig. 7. A flow chart of the system characterization process.

The first form is the first-order lag-plus-dead-time model. This has the Laplace domain form

$$F(S) = K\,[\exp(-ST_D)]\,/(\tau S + 1) \tag{5}$$

where K is the system gain, T_D is the delay time, and τ is the time constant.

For an input step in enzyme concentration from 3 to 12 IU l^{-1}, i.e. a total change of 9 IU l^{-1}, the step-response time-domain expressions are

$$\begin{aligned} r(t) &= r_0 && \text{for all } t \leqslant T_D \\ r(t) &= r_0 + 9K\,\{1 - \exp[-(t - T_D)/\tau]\} && \text{for all } t \geqslant T_D \end{aligned} \tag{6}$$

where r_0 is the initial value of $r(t)$ at $t = 0$. A non-linear least-squares technique [7, 8] was used to determine the parameters of eqn. (6). The results were: $K = 1.42$ μmol/IU, $T_D = 66.8$ s and $\tau = 26.7$ s. The actual fit to the experimental data is shown in Fig. 8A. From the standpoint of designing a control system, this fit is adequate.

To improve matters, a second-order lag-plus-dead-time model having two independent time constants can be used. This has the Laplace domain form

$$F(S) = K\,[\exp(-ST_D)]\,/(\tau_1 S + 1)(\tau_2 S + 1) \tag{7}$$

where K is the system gain, T_D is the delay time, τ_1 is the first time constant and τ_2 the second time constant.

The step-response time-domain expressions are

$$r(t) = r_0 \quad \text{for all } t \leqslant T_D$$

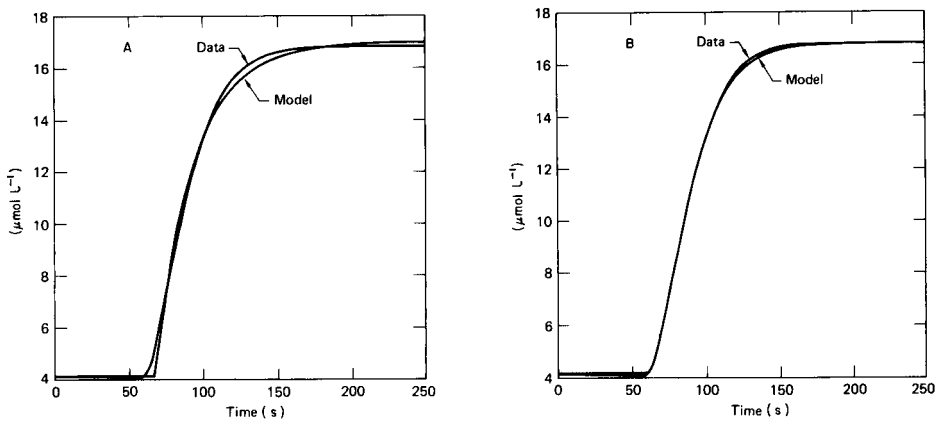


Fig. 8. Comparison of the model and actual data for (A) a first-order model, and (B) a second-order model.

$$r(t) = r_0 + 9K\{1 + \tau_1 \exp[-(t - T_D)/\tau_1] - \tau_2 \exp[-(t - T_D)/\tau_2]\}/(\tau_2 - \tau_1)$$

$$\text{for all } t \geqslant T_D \qquad (8)$$

Again the non-linear least-squares technique was used to determine the parameters of $r(t)$. The results were: $K = 1.41$ μmol/IU, $T_D = 59.6$ s, $\tau_1 = 16.0$ s, and $\tau_2 = 16.0$ s.

The fit (Fig. 8B) is better than that of the first-order lag. Because the two time constants converge to equal values, a somewhat simpler model is possible:

$$F(S) = K[\exp(-ST_D)]/(\tau S + 1)^2 \qquad (9)$$

For this model, the step-response time-domain expressions are

$$r(t) = r_0 \qquad \text{for all } t < T_p$$

$$r(t) = r_0 + 9K\{1 - [1 + (t - T_D)/\tau] \exp[-(t - T_D)/\tau]\} \quad \text{for all } t \geqslant T_D \qquad (10)$$

The non-linear least-squares technique yields the parameters: $K = 1.41$ μmol/IU, $T_D = 59.6$ s and $\tau = 16.0$ s. The fit is essentially the same as before, but only three parameters are necessary to describe the system dynamics. This fit can be used to calculate the convolution function shown in Fig. 9.

CONCLUSIONS

Any approach to modeling a system must consider the complexity of the system and the purpose of the model. For some systems it is possible to start from "first principles" and derive an analytical model. When the problem is tractable, this is usually the best choice. As system complexity increases,

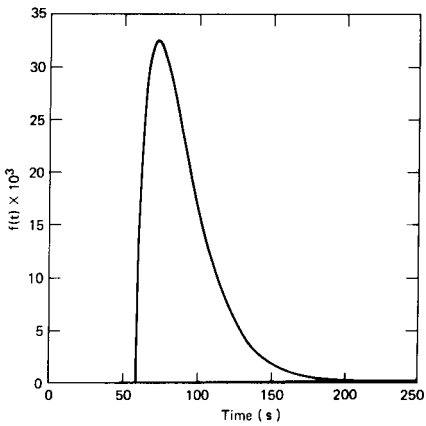


Fig. 9. Convolution function computed for the second-order model.

this approach becomes unreasonable if not impossible. An alternative method is to observe the data and propose an analytical model that represents the dominant properties of the system. When a single model fails to provide satisfactory results, multiple local models can be combined to form an improved model of the entire system. If the level of system complexity becomes even greater, it may be desirable to use a numerical data model derived from a well-controlled experiment as a representation of the system.

The analytical model and the numerical model developed above for representing the flow dynamics of an enzyme-catalyzed reaction can both be considered satisfactory. However, the analytical model is a more compact representation than the data model. A second-order lag-plus-dead-time model was shown to be a good fit to the data. Once an analytical model has been developed, the apparatus can be used for the study and control of dynamic or steady-state chemical processes.

This work was performed under the auspices of the U.S. Department of Energy under contract No. W-7405-Eng-48. The authors acknowledge helpful discussions with Lester P. Rigdon and David J. Balaban.

REFERENCES

1 J. W. Frazer, L. P. Rigdon, H. R. Brand and C. L. Pomernacki, Report UCRL-80765 (1978). Anal. Chem., accepted for publication.
2 J. W. Frazer, L. P. Rigdon, H. R. Brand, C. L. Pomernacki and T. A. Brubaker, Report UCRL-80139 (1978). Anal. Chem., accepted for publication.
3 P. Eykhoff, System Identification, Wiley, New York, 1974.
4 G. Cooper and C. McGillen, Continuous and Discrete Signal and Systems Analysis, Holt Rinehart and Winston, New York, 1974.
5 A. Savitzky and M. J. E. Golay, Anal. Chem., 36 (1964) 1627.
6 D. D. Perlmutter, Introduction to Chemical Process Control, Wiley, New York, 1965, p. 140.
7 Program ZXSSQ, (Levenberg—Marquart algorithm), International Mathematical and Statistical Libraries, Inc., Houston, TX.
8 Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974.

Analytica Chimica Acta, 112 (1979) 297—312
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

MULTIPLE-SCANNING POTENTIOMETRIC STRIPPING ANALYSIS

JOHN MORTENSEN, EZRA OUZIEL**, HANS JØRGEN SKOV and LARS KRYGER*
Department of Chemistry, University of Aarhus, Langelandsgade 140, DK-8000 Aarhus C (Denmark)

(Received 2nd March 1979)

SUMMARY

A computerized data acquisition technique — multichannel potentiometric monitoring — is used in conjunction with potentiometric stripping analysis. Multiple-scanning stripping potentiograms can be recorded so that the analytical signals are enhanced. A minicomputer with an internal store of 4K 16-bit words suffices for experimental control and data treatment. The technique is suitable for stripping analysis with preconcentration times of 60—90 s at a mercury film electrode with linear response ranges of 1—100 $\mu g\ l^{-1}$ for cadmium(II) and lead(II). For preconcentration times of 30 min, the limit of detection is about 5 $ng\ l^{-1}$. Preliminary tests on continuous flow analysis are reported.

Electrochemical stripping analysis is a highly sensitive technique for the determination of various trace substances in solution [1]. Conventional anodic stripping voltammetry is applicable to several heavy metals and cathodic stripping voltammetry has proved suitable in analyses for halides and sulphides. Several techniques have been employed for stripping analysis. These include d.c. and a.c. voltammetric, pulsed voltammetric and potentiometric methods. This paper deals with a new computerized multiple-scanning technique of stripping analysis.

Conventional voltammetric stripping analysis for trace metals at or below $\mu g\ l^{-1}$ concentrations mostly requires long plating times (of the order of hours). As sample contamination and a change in sample composition may increase with increasing plating times, it is desirable to develop techniques which reduce the plating times.

In certain types of trace metal analysis, lengthy ashing or extraction procedures are necessary prior to the actual instrumental analysis. In such cases a plating time of a few hours is of little importance. Many types of liquid samples may, however, be directly subjected to stripping analysis. In such cases the time per sample will be almost proportional to the plating time. Moreover, the increasing demand for rapid analysis of a multitude of

**Present address: Technion, Haifa, Israel. Permanent address: Abic Ltd., Chemical and Pharmaceutical Industries, P.O.B. 2077, Ramat-Gan, Israel.

very small samples carried by a continuously flowing stream stresses the general need for techniques with short turn-round times.

The technique described here is based on the potentiometric stripping method reported by Jagner et al. [2—4]. Apart from the basic instrumentation used by these authors, a small digital computer in conjunction with an interface module for controlling the experiment, and acquiring and processing the data, is included. This modification makes signal enhancement by multiple scanning possible: after a single short preconcentration period, a considerable fraction of the accumulated metal may be forced to undergo several oxidation and re-reductions in a precisely timed sequence. The computer acquires and adds the analytical signals resulting from the oxidation steps. Thus even after a short plating period, a relatively small amount of preconcentrated metal may produce a significant analytical signal.

As is the case in stripping voltammetry, potentiometric stripping analysis involves an electrolytic preplating of dissolved trace metals into a mercury working electrode. In potentiometric stripping analysis, however, potentiostatic control is maintained only during the plating period. Unlike voltammetric stripping analysis, where a linear or modified linear potential scan gives rise to the redissolution of amalgamated metals, this potentiometric technique employs a chemical oxidation of the preconcentrated metals by some oxidizing agent, e.g. mercury(II), in solution. The redissolution process closely resembles a redox titration with constant delivery of titrant. The potential versus time behaviour of the working electrode is monitored and the lengths of the characteristic plateaux, each of which indicates the oxidation of a metal, are taken as a measure of the concentrations.

PRINCIPLES

The feasibility of signal enhancement by the application of computerized multiple scanning to voltammetric stripping analysis has been demonstrated [5]. As some of the principles utilized are likely to apply to potentiometric stripping analysis, some essential observations are summarized as follows.

In multiple-scanning voltammetric stripping [5], enhancement of the analytical signal is strongly dependent on the recovery (re-reduction) of newly oxidized metals after each cyclic scan. If the time available for oxidized metals to escape from the working electrode by diffusion in a quiet solution is short, the recovery is large. Thus optimum signal enhancement is obtained by employing fast anodic sweeps; the oxidation potential is scanned only as far as is required to obtain the signal and this is followed by a prompt return to the reduction potential. Similarly, in multiple-scanning potentiometric stripping analysis, chemical oxidation should proceed quickly followed by resumption of potentiostatic control at the reduction potential. As in potentiometric stripping analysis, the rate of the oxidation process may be controlled by the amount of oxidizing agent added to the solution; a large recovery of metals should result if a large excess of oxidizing agent is employed.

Multichannel potentiometric monitoring

Figure 1(I) shows schematically the potential versus time behaviour of a mercury film electrode during potentiometric stripping of three amalgamated species, the length of the plateaux being proportional to the metal concentrations in the film and, for a given pre-electrolysis time, in the sample. In order to add two or more such potentiograms, these must be given as time versus potential functions sampled at regular potential intervals. Indeed, straightforward monitoring of potential versus time at regular time intervals can be employed. However, as the discrete set of potentials obtained during the monitoring of a particular potentiogram does not necessarily coincide with any other set of potentials obtained during other scans, even when the discrete set of times match, addition of potentiograms is possible only after some interpolation procedure, which in turn introduces an uncertainty.

The most obvious way of recording the time versus potential behaviour of the working electrode sampled at regular potential intervals is indicated in Fig. 1(II, III). Initially, a data buffer within the computer memory is defined and reset to all zeroes. A correspondence between a potential reading E and the address A of a word in the computer memory is established in the following manner:

$$A = A_0 + \text{ROUND}\ (L^* (E - E_c)/(E_a - E_c))$$

where A_0 defines the first buffer address, L the buffer length in words, and E_c and E_a define the cathodic and anodic limits of the potential range

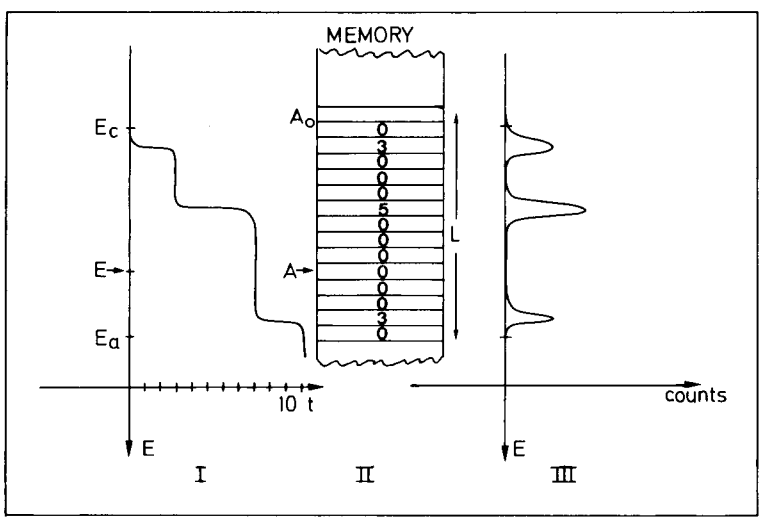


Fig. 1. (I) Potential versus time behaviour of working electrode during redissolution of three amalgamated metals. The potential window E_c, E_a is studied. (II) Computer memory section: the data storage area starting at address A_0 holds a record of accumulated clock pulse counts. (III) The resultant multichannel potentiogram.

("window") selected for study. At the end of the pre-electrolysis in a stirred solution, the computer stops the stirring and after 30 s potentiostatic control is abandoned. The potentiogram is then sampled by continuously monitoring the electrode potential during the redissolution period and by using a programmable clock to determine the sampling rate. If a potential reading lies within the selected window, the corresponding memory location is incremented by 1. If $E < E_c$ the point is omitted; if $E > E_a$ data acquisition is terminated. In multiple-scanning potentiometric stripping the electrolysis potential is then immediately resumed for a short period and the procedure repeated, without clearing the data buffer. Figure 1(II) illustrates the clock counts accumulated in the computer memory. The resulting potential distribution function is shown in Fig. 1(III). Clearly a plateau in the conventional potentiogram corresponds to a peak in the distribution function; the area of the peak is proportional to the length of the plateau, so that quantitative results can be obtained by peak integration, which is easily done by the computer. The conventional potentiogram may if desired be obtained by integration of the entire function.

The mode of data acquisition resembles that of a multichannel analyser in monitoring the energy distribution of radiation quanta and may be described as multichannel potentiometric monitoring.

EXPERIMENTAL

Interface module, computer and software

A general-purpose computer-controlled interface module for analytical work was used. For testing of the electroanalytical procedures, the unit provides the facility of generating linear potential ramps, cyclic potential ramps and potential steps. Potential ramps may be modulated by rectangular pulse trains of variable amplitudes and frequencies. Automatic switching between potentiostatic control with cell-current monitoring and a potentiometric mode with cell-potential monitoring is provided. Twelve 220-V a.c. on—off switches and three digital-to-analog (D/A) converters are included. Analog signals are acquired by means of a 12-bit A/D converter (conversion time 20 μs) triggered by a programmable quartz-clock pulse generator.

The computer employed is an ALPHA LSI-3/05 naked version with an internal store of 4K 16-bit words. This minicomputer is connected to a departmental time-shared RC4000 computer via a common teletype line. For the present project this communication line was used solely for loading the ALPHA with programs assembled in the RC4000 [6]. Once loaded, the ALPHA acts as an independent computer: communication with the experimenter is via an ADM-3A alphanumeric screen. Acquired data may be inspected on a dynamic graphic display [7] or (via two of the D/A converter drawn by a Philips PM 8041 X—Y recorder. These peripheral devices are connected to the ALPHA and are directly available to the experimenter.

The computer program is written in the RC4000—ALPHA LSI-3/05 cross

assembler language. The entire program including the analytical procedure, driver routines for the electroanalytical module, and the displays and data storage areas occupies approximately 2K words.

Further details of the hard-ware and soft-ware will be published later [8].

Electrochemical cell, glassware, flow cell and pump

The electrochemical cell was a glass vessel mounted in a titration assembly (Radiometer TTA60). The sample solution could be stirred with a PTFE coated stirrer. Samples were deaerated by passing argon purified in an active charcoal filter through the solution for 15 min prior to analysis. During analysis, argon was passed over the sample.

Prior to analysis all glassware was thoroughly cleaned in (1 + 1) nitric acid and washed with the sample solution several times.

A flow cell of the wall-jet type [9] was used for monitoring trace metals in continuously flowing stream. The peristaltic pump used was an Ismatec Mini-S model.

Electrodes

A glassy carbon rod embedded in PTFE (Radiometer F3500) plated with mercury was used as the working electrode. The reference electrode was a saturated calomel electrode (Radiometer K4040) and a platinum wire (5 mm long, 0.5 mm thick) served as the counter electrode.

In order to obtain a thin mercury film, the working electrode was held at 0.0 V vs. SCE in a stirred deaerated solution of mercury(II) acetate (200 mg $Hg\,l^{-1}$) in 0.05 M hydrochloric acid for 1 min. Then the potential was decreased to −0.5 V vs. SCE and kept at this value for 2 min. As the film thus formed is very stable, the electrode was cleaned and replated only once per week. About once every two weeks the electrode surface was polished with 0.25-μm diamond paste and replated.

Chemicals

Doubly-distilled water and analytical-grade chemicals were employed for most experiments. However, to obtain quantitative results for analytes at concentrations below 10 $\mu g\,l^{-1}$, mineral acids and sodium chloride of Suprapur grade (Merck) were used.

RESULTS AND DISCUSSION

In order to elucidate the characteristics of multiple-scanning potentiometric stripping analysis, experimental parameters were varied and their effects on the recorded potentiograms were studied. Unless explicitly stated otherwise, these experiments involved a preconcentration period of 60 s at −1300 mV vs. SCE in a stirred solution followed by a period of 30 s without stirring. The data acquisition rate was 1 point per 300 μs. The samples studied contained 1 mg $Cd\,l^{-1}$ (as cadmium chloride) and 1 mg $Pb\,l^{-1}$ (as lead nitrate) in 2×10^{-3} M hydrochloric acid.

Effect of oxidant concentration

In order to study the effect of the mercury(II) (oxidant) concentration on the recovery of metals as a function of the number of scans, 1, 5 and 10 ml of a mercury(II) solution (4 g l^{-1}) were added to 50-ml samples, and the individual potentiograms obtained during the 1st, 2nd, 4th, 16th, 64th and 128th scan were recorded. The results are shown in Fig. 2: obviously, the amplitude of the signal generally tends to decrease with increasing scan number, the decrement being more severe for cadmium than for lead. Nevertheless, a significant signal remains for all mercury(II) concentrations even after 128 scans. Although, as pointed out by Jagner and Graneli [2], in potentiometric stripping analysis the total amount of amalgamated metals is removed from the electrode by oxidation, repetitive scanning is a profitable process in the sense that the material oxidized may be partly recovered and redetermined. Figure 2 also shows that an increasing mercury(II) concentration results in lower amplitudes. However, the relative decrement of

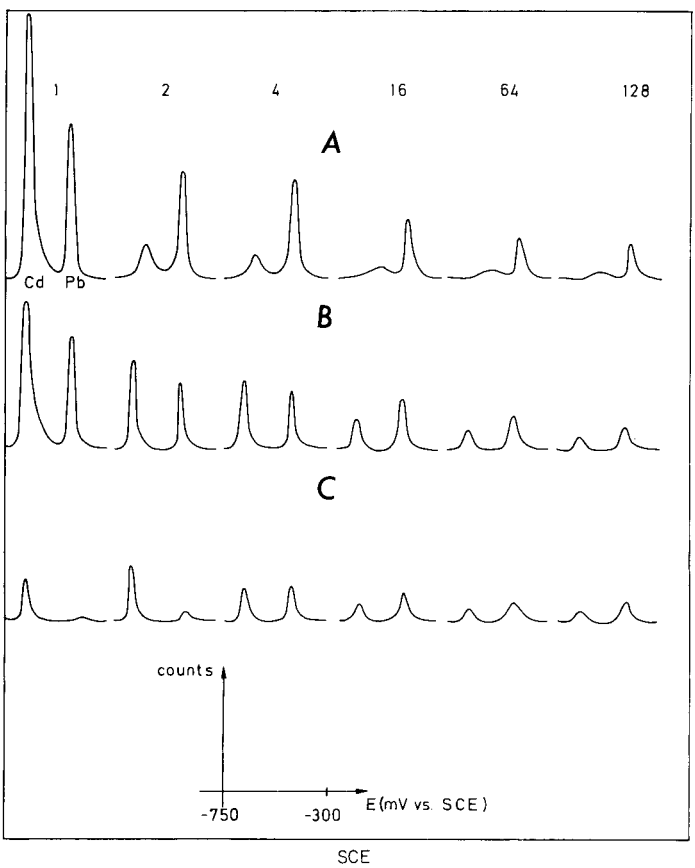


Fig. 2. Signal dependence on scan number and concentration of oxidizing agent. Individual potentiograms shown after 1, 2, 4, 16, 64 and 128 scans. Mercury(II) concentrations: (A) 78 mg l^{-1}; (B) 364 mg l^{-1}; (C) 667 mg l^{-1}.

the cadmium peak becomes less severe. This behaviour may be attributed to the fact that rapid stripping of lead leaves a shorter time for the recently oxidized amount of cadmium to escape from the vicinity of the electrode. For very high concentrations of mercury(II) (Fig. 2C) the lead peak increases during the first 4 scans. This behaviour is not fully understood, but may be related to the formation of a thick mercury film and the presence of an excessive amount of oxidizing agent in the immediate vicinity of the film.

Data acquisition rate

Apart from being a function of the concentrations of analytes and oxidizing agent, the signal amplitude increases with the data acquisition rate, i.e. the clock pulse rate. Figure 3 shows a potentiogram recorded at two different sampling rates. Obviously, the signal-to-noise ratio may be improved by increasing the clock frequency. The maximum data acquisition rate with the present system is 1 datum per 300 μs because of extensive software checks. However, with a system exclusively designed for multiple-scanning potentiometry the frequency could be increased by an order of magnitude and hence better "counting statistics" achieved.

Effect of moving the potential window

To 50 ml of the sample was added 0.5 ml of mercury(II) solution (4 g l^{-1}). A series of analyses was done; in each case the results of 16 scans were added. The potential window was moved so that the anodic limit was decreased by 37.5 mV between consecutive experiments. The results are shown in Fig. 4. Here a marked increment of the cadmium peak is observed, when the time spent at potentials where lead is oxidized, is decreased.

Reduced pre-electrolysis times

In single-scan potentiometric stripping analysis, the analytical signal, i.e. the length of the potential plateau observed during the stripping of a given

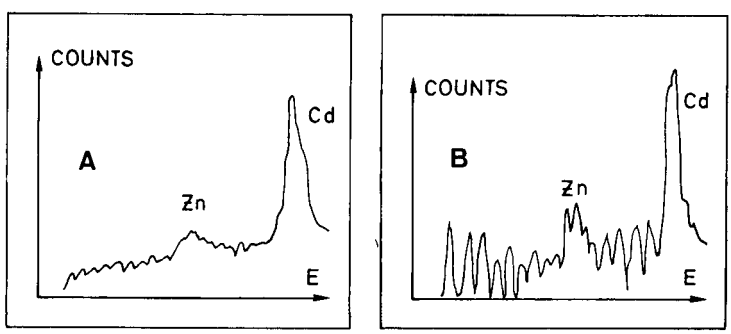


Fig. 3. Signal-to-noise ratio dependence on sampling rate of a 20 μg Cd l^{-1} solution electrolysed for 20 s. Data sampling rate: (A) fast, 750 μs/datum; (B) slow, 3000 μs/datum.

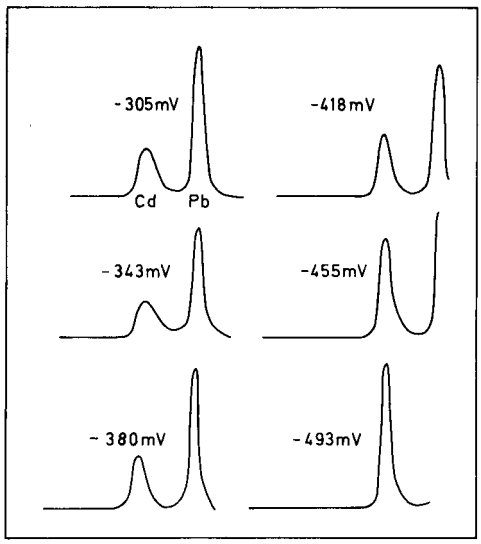


Fig. 4. Dependence of cadmium signal on the anodic limit of the potential window.

amount of amalgamated metal, may be increased by (a) extending the pre-concentration period, (b) decreasing the concentration of oxidizing agent, or (c) performing the stripping process in a quiet rather than in a stirred solution [2, 3]. Therefore, theoretically, it should be possible to determine metals in the $\mu g\ l^{-1}$ concentration range by single-scan stripping at very short electrolysis times, simply by lowering the concentration of oxidizing agent sufficiently. However, as pointed out by Jagner and Graneli [2], the concentration of oxidizing agent should be well in excess of the total concentration of analytes in order to lower the effect of analytes in solution acting as oxidizing agents themselves, e.g. lead(II) oxidizing amalgamated cadmium. Furthermore, traces of dissolved oxygen always impose a practical lower limit of oxidizing agent. Thus for low concentrations of analytes, prolonged electrolysis times and possibly stripping in a quiet solution are necessary to obtain a detectable signal in single-scan stripping.

In order to evaluate the possibilities of reducing pre-electrolysis periods by employing the multiple-scanning technique, electrolysis was carried out in 50 ml of 5×10^{-3} M hydrochloric acid containing 20 mg Hg(II) l^{-1}. The electrolysis was carried out at —950 mV vs. SCE in a stirred solution for 1 min followed by a period of 30 s without stirring. The subsequent single-scan potentiometric stripping curve is shown in Fig. 5 (curve A); curve B is the potentiogram obtained from a similar experiment after the addition of cadmium(II) ($2\ \mu g\ l^{-1}$). The potential plateau obtained lasts for ca. 0.2 s, which is just long enough to be resolved by a X—T recorder in a non-computerized approach. Curves C and D show the potentiograms obtained from this spiked solution by monitoring 1 and 16 scans, respectively. Clearly,

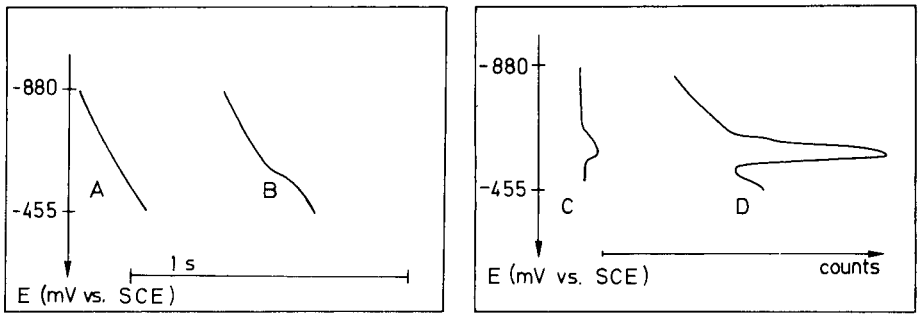


Fig. 5. Comparison between single-scanning and multiple-scanning potentiometric stripping analysis of cadmium(II) (2 $\mu g\ l^{-1}$) preconcentrated for 60 s at —950 mV vs. SCE. (A) Single scan, blank solution; (B) single scan, spiked solution; (C) multichannel potentiogram of spiked solution, 1 scan; (D) multiple-scanning potentiogram, 16 scans.

the multiple-scanning technique results in a significant signal even for a short preconcentration time.

Quantitative analysis and interrupted stripping

To obtain reliable results in trace metal analysis, the standard addition method is frequently preferred. The applicability of this method to multiple-scanning potentiometric stripping analysis was tested by preparing a calibration curve for cadmium in the range 1—13 $\mu g\ l^{-1}$. For each standard addition, preconcentration was done at —1300 mV vs. SCE for 1 min, followed by 30 s in an unstirred solution. The potential window used during the stripping was set so that the anodic limit was —518 mV vs. SCE, i.e. slightly anodic of the cadmium peak and cathodic of the lead peak. A total of 16 scans was recorded and added. A linear relationship between the added amount of cadmium and the observed signal was obtained.

When, however, multi-element analysis was attempted by scanning both the cadmium and the lead peak, the plot for lead was linear, whereas the calibration plot for cadmium was markedly curved. The latter non-linear behaviour may be attributed to the fact that in multiple-scanning potentiometric stripping the recovery of an element after each scan depends strongly on the time used for the stripping of other elements at more anodic potentials. Thus, the recovery of cadmium can be expected to be diminished as the lead concentration increases. This situation is not encountered in multiple-scanning voltammetric stripping as the potential scan time does not depend on the concentration of analytes.

In order to avoid this dependence, the analytical procedure was modified so that, after the preconcentration period, the amalgamated metals were stripped in a two-step process: interrupted stripping. First, the anodic limit of the potential window was set between the stripping potentials of cadmium

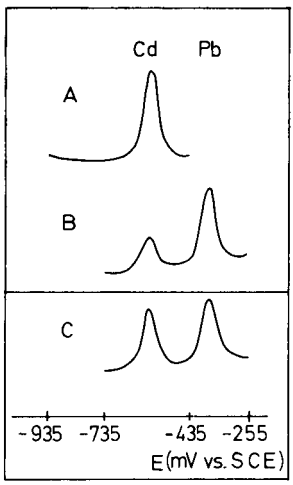


Fig. 6. Interrupted potentiometric stripping of cadmium(II) and lead(II). A total of 16 scans is added. (A) Stripping interrupted after cadmium(II) redissolution; (B) stripping interrupted after lead(II) redissolution; (C) uninterrupted stripping of both metals.

and lead, and 16 scans were performed. The result is shown in Fig. 6 (curve A). As after each scan the working electrode was immediately reset to the electrolysis potential, lead was not stripped during these scans. Then the potential window was moved so as to include the lead peak, and 16 more scans were performed (curve B). It should be pointed out that curves A and B were obtained in rapid succession, i.e. after only one pre-electrolysis period. Curve C shows the result obtained by using the uninterrupted procedure. Clearly, both the cadmium and lead peaks are increased when the interrupted procedure is applied. The cadmium peak is increased because of increased recovery between scans. The lead peak is slightly increased because during the 16 scans of the cadmium peak, lead(II) is still being continuously reduced. This prolonged electrolysis of lead implies that the lead signal depends on the cadmium concentration. The dependence vanishes when the amount of lead deposited during the preconcentration step is large compared to the amount deposited during the cadmium redissolution. More effective stirring during the preconcentration and subsequent stripping in a quiet solution were, therefore, used in the following experiments.

Figure 7 illustrates the advantage of multiple-scanning potentiometric stripping in the interrupted stripping mode, which removes the dependence between the cadmium and lead signals. When this procedure was used, linear standard addition plots were obtained for both metals in the range 1—100 $\mu g\,l^{-1}$ (Fig. 8). For the interrupted stripping procedure E_a = —455 mV vs. SCE was used for the cadmium determination, and E_a = —255 mV vs. SCE for the lead determination.

Sample dilution in the standard addition method. Whenever aliquots of standard solutions are added to a sample, the analyte concentration is

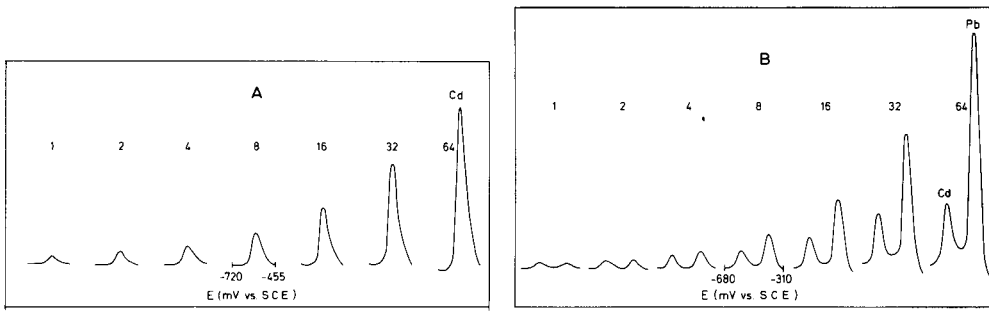


Fig. 7. Signal enhancement for (A) cadmium and (B) lead by multiple-scanning potentiometric stripping analysis with the interrupted stripping technique. Sample: 20 μg Cd l^{-1} and 20 μg Pb l^{-1} in 0.04 M hydrochloric acid; 160 mg Hg l^{-1}; pre-electrolysis at −1300 mV vs. SCE for 60 s. 1, 2, 4, 8, 16, 32 and 64 scans are added.

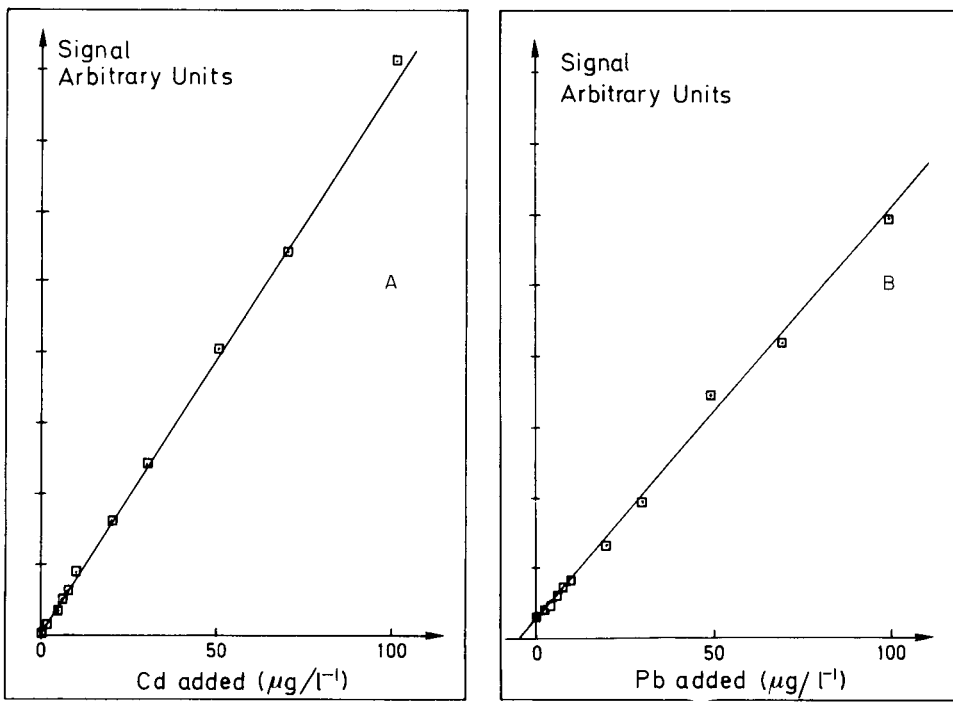


Fig. 8. Standard addition analysis of (A) cadmium(II) and (B) lead(II): 50 ml of solution containing 0.1 M hydrochloric acid, 0.5 M sodium chloride and 400 mg Hg l^{-1}. Pre-electrolysis for 60 s at −1300 mV vs. SCE; 64 scans are added.

diminished by dilution. In stripping voltammetry, it is common practice to correct currents accordingly. In potentiometric stripping analysis, however, both the analyte and the oxidizing agent are diluted by the same fraction, and no correction is necessary (unless of course the oxidizing agent has been added to both standard and sample).

Evaluation of analytical signal

Although in general the background capacitance signal is low in potentiometric stripping analysis [3], its contribution becomes important when the ratio of amalgamated metal concentration to the concentration of oxidizing agent is high. In multiple-scanning potentiometric stripping high concentrations of oxidizing agents are generally employed, and the background is negligible. The background may if desired be corrected for by subtracting from the potentiogram a "blank" potentiogram obtained by carrying out the analysis without the preconcentration step. Figure 9 shows the total signal and the background signal from the analysis of a solution containing mercury(II) and spiked with cadmium(II) and lead(II). After pre-electrolysis in a stirred solution for 60 s and a period of 30 s to allow the solution to become still, 16 scans with E_a = −330 mV were added by using the uninterrupted mode. The solution was then stirred for 10 s and allowed to rest for 30 s, and the experiment was repeated without the plating. Obviously, the background is very small, though traces of cadmium and lead are visible (Fig. 9, curve b). The reason for this is that the plating time can never be exactly zero; during the 16 background scans some metal will be accumulated.

As the cadmium and lead peaks overlap to some extent, the heights of the peaks rather than the area should be used as a measure of concentration. The validity of this mode of signal evaluation was confirmed by the standard addition plots obtained.

Sub-part per billion concentrations

To investigate whether rapid analysis of trace metals in the sub-ppb concentration range was possible, the standard addition method was applied to 50 ml of 0.5 M sodium chloride containing 0.06 M hydrochloric acid and

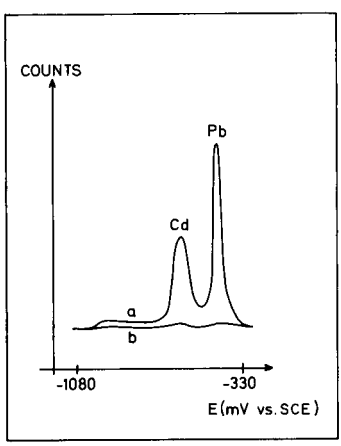


Fig. 9. Multiple-scanning stripping potentiograms of 50 ml of 0.01 M hydrochloric acid containing 40 mg Hg l^{-1}, 50 μg Cd l^{-1} and 50 μg Pb l^{-1}; 16 scans are added. (A) Pre-electrolysis at −1300 mV for 60 s; (B) without pre-electrolysis.

mercury(II) (240 mg l^{-1}). The sample was plated for 60 s in a stirred solution at −1300 mV; after a further 30 s, 64 scans with E_a = −485 mV vs. SCE were accumulated. As the concentration of lead(II) in the unspiked sample was of the order of 5 μg l^{-1}, only cadmium(II) was considered; for each standard addition, duplicate analyses were done. The results (Fig. 10) indicate that the background concentration of cadmium(II) is ca. 0.1 μg l^{-1}.

Precision, detection limit and accuracy

The precision obtained by multiple-scanning potentiometric stripping analysis depends on many parameters, e.g., the concentrations of analytes and oxidizing agent, preconcentration time and data sampling frequency. To estimate the precision, a 50-ml sample of 0.5 M sodium chloride containing 0.04 M hydrochloric acid and 160 mg Hg l^{-1} was spiked with cadmium(II) and lead(II) (each 20 μg l^{-1}). Consecutive analyses (10) were done at a plating potential of −1300 mV vs. SCE for 60 s in a stirred solution, and after 30 s, 64 scans were accumulated in the interrupted mode (E_a = −455 and −310 mV vs. SCE for cadmium and lead, respectively). The standard deviations found were 0.018 for cadmium and 0.024 for lead.

The detection limit depends on the preconcentration time. For solutions spiked with cadmium and lead, the analytical signals varied proportionally with the pre-electrolysis time from 1 to 30 min. If 30 min is regarded as the maximum acceptable time, the detection limits for cadmium and lead are of the order of 5 ng l^{-1}.

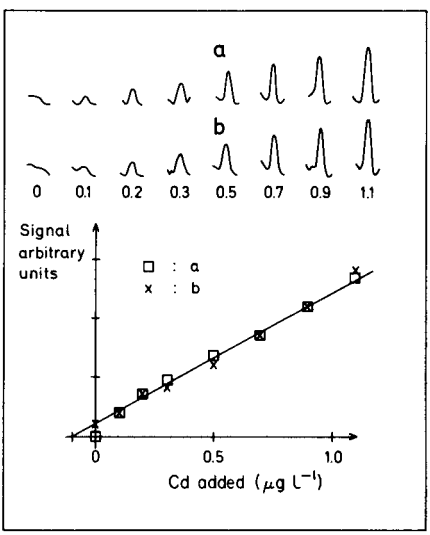


Fig. 10. Plot obtained by the standard addition method for 50 ml of 0.5 M sodium chloride containing 0.06 M hydrochloric acid and 240 mg Hg l^{-1}. Pre-electrolysis for 60 s at −1300 mV vs. SCE; 64 scans added; (a, b) signals as solution is spiked with 0, 0.1, 0.2, 0.3, 0.5, 0,7, 0.9 and 1.1 μg Cd l^{-1} (duplicate runs).

To estimate the accuracy of the proposed method the contents of cadmium and lead in a standard biological material (Bowen's kale [10]) were determined. Samples (0.2 g) of the kale were digested with 3 ml of concentrated nitric acid in a PTFE-lined pressure bomb at 150°C overnight; 0.5 ml of concentrated hydrochloric acid was then added to each sample and the volumes were made up to 50 ml with triply-distilled water. The results obtained are shown in Table 1, along with the recommended best mean values and ranges [10].

Detection of interferences

Metals with similar stripping potentials will obviously interfere with each other. Such interferences may often be resolved by a proper choice of solvent system. The multiple-scanning potentiometric stripping technique may be helpful in detecting the presence of an interfering element. First, as it produces essentially the derivative of the conventional stripping potentiogram it is more detailed; secondly, repetitive scanning (uninterrupted mode) tends to enhance the signal of the components with anodic redissolution potentials. Figure 11 (curve A) shows the conventional single-scan stripping potentiogram for a mixture of cadmium(II) and thallium(III) in acetate medium. This potentiogram indicates that cadmium is present with a zinc impurity, but thallium is barely discernible. Curve B shows the corresponding multiple-scan potentiogram after the addition of 256 scans; although badly resolved, zinc, thallium and a trace of lead have become visible.

Possible future applications

Continuous flow stripping analysis. As multiple-scanning potentiometric stripping analysis is characterized by short electrolysis periods, it is well suited for rapid trace analysis of large numbers of samples. To evaluate the possibilities of monitoring trace metals in a flowing stream, a simple electrochemical flow detector of the wall-jet type [9] was constructed and a solution containing 0.5 M sodium chloride, 0.02 M hydrochloric acid, and 80 mg $Hg\,l^{-1}$ was pumped through the cell at a flow rate of 2 ml min^{-1}. This solution was spiked with 40, 80 and 120 $\mu g\ l^{-1}$ concentrations of cadmium(II) and lead(II)

TABLE 1

Comparison of results obtained by the proposed method for standard kale with the best mean values and ranges reported by Bowen [10]

Sample	Proposed method ($\mu g\ g^{-1}$)		Best mean value (range) ($\mu g\ g^{-1}$)	
	Cd	Pb	Cd	Pb
1	0.77	2.20		
2	0.92	1.80	0.80 (0.38—1.06)	2.645 (1.6—3.8)
3	0.87	1.88		
Mean	0.85	1.96		

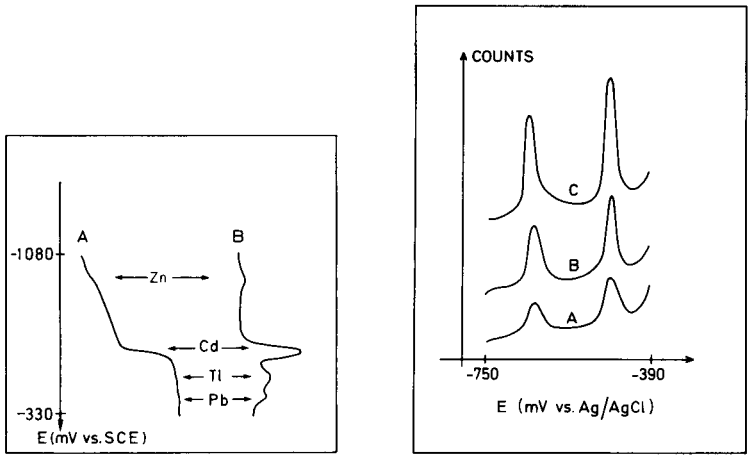


Fig. 11. Thallium(III)—cadmium(II) interference in potentiometric stripping analysis of 50 ml of 0.1 M acetic acid—0.1 M sodium acetate containing 0.2 g Hg l^{-1}, 40 μg Cd l^{-1} and 40 μg Tl l^{-1} with pre-electrolysis at −1300 mV vs. SCE for 1 min. (A) Single-scan stripping potentiogram; (B) multiple-scanning stripping potentiogram after 256 scans.

Fig. 12. Continuous-flow stripping analysis of 500 μl of a solution containing 0.5 M sodium chloride, 0.02 M hydrochloric acid and 80 mg Hg l^{-1} at a flow rate of 2 ml min^{-1}, spiked with cadmium(II) and lead(II): (A) 40 μg l^{-1}; (B) 80 μg l^{-1}; (C) 120 μg l^{-1}.

For each concentration, multiple-scanning stripping potentiograms were obtained by electrolysing at −1300 mV vs. SCE for 5 s followed by immediate acquisition and addition of 16 scans. The liquid flow rate remained constant during the analysis, but even so there was good recovery between the scans. Figure 12 shows the resulting potentiograms. The total time required to obtain one of these potentiograms was 14 s, including plating and stripping. During this time less than 500 μl of solution passes through the detector. The technique should, therefore, be suitable for the analysis of samples injected into a continuously flowing stream [11, 12].

Samples with high oxidant concentrations. Throughout this work mercury(II) was used as the oxidant. However, the oxygen present in non-deaerated samples may be used to oxidize several of the less noble elements [13]. This approach is advantageous as pretreatment of the sample with purging gas and mercury(II) is unnecessary. For low analyte concentrations and short plating times, the ratio of the concentrations of oxidant and analyte is unfortunately so high that the subsequent rapid stripping cannot be observed on a normal X—T recorder. However, if the computerized technique described above is employed, the advantages of analysing non-deaerated samples may be better exploited.

The support of The Danish Natural Science Research Council (grant no. 511-8032) is gratefully acknowledged. The authors thank Dr. H. J. M. Bowen, University of Reading, England, for the donation of the standard kale.

REFERENCES

1 F. Vydra, K. Stulik and E. Julaková, Electrochemical Stripping Analysis, Ellis Horwood, Chichester, 1976.
2 D. Jagner and A. Graneli, Anal. Chim. Acta, 83 (1976) 19.
3 D. Jagner and K. Årén, Anal. Chim. Acta, 100 (1978) 375.
4 D. Jagner, Anal. Chem., 50 (1978) 1924.
5 L. Kryger and D. Jagner, Anal. Chim. Acta, 78 (1975) 250.
6 J. Nyborg, B. Nielsen and T. Bogh, RC4000 ALPHA Cross Assembler, University of Aarhus, 1977.
7 H. J. Skov, L. Kryger and D. Jagner, Anal. Chem., 48 (1976) 933.
8 H. J. Skov and L. Kryger, Anal. Chim. Acta, to be published.
9 B. Fleet and C. J. Little, J. Chromatogr. Sci., 12 (1974) 747.
10 H. J. M. Bowen, J. Radioanal. Chem., 19 (1974) 215.
11 J. Ružička and E. H. Hansen, Anal. Chim. Acta, 99 (1978) 37.
12 H. Mosbæk and J. Hansen, in preparation.
13 D. Jagner, Anal. Chem., 51 (1979) 342.

Analytica Chimica Acta,112 (1979) 313—319
Computer Techniques and Optimization
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

ERMITTLUNG DER ANALYSENFREQUENZ BEI DISKONTINUIERLICHEN PROZESSANALYTISCHEN VERFAHREN

KLAUS DOERFFEL* und GERALD LORENZ

Technische Hochschule "Carl Schorlemmer", Leuna-Merseburg, 42 Merseburg (Deutsche Demokratische Republik)

ILONA TAGLE

Nationales Forschungszentrum (CNIC), Havanna (Cuba)

(Eingegangen am 7 Februar 1979)

ZUSAMMENFASSUNG

Bei Anwendung des Abtasttheorems auf Prozesse der stoffwandelnden Industrie ergeben sich durch die Eigenarten der Prozesse und des analytischen Arbeitens Besonderheiten. Neben den Instationaritäten der Prozesse sind die endliche Länge der Signalaufzeichung und der damit verbundene statistische Fehler der Korrelationsfunktion zu berücksichtigen. Bei Kenntnis der Korrelationsdauer des stochastischen Signals kann man den zeitlichen Abstand zweier Analysen in Abhängigkeit von der Prozeßstreuung und der zulässigen analytischen Toleranz angeben.

SUMMARY

Establishment of the frequency of analysis for discontinuous measurements in process control.

Application of sampling theory to industrial chemical processes involves special problems arising from the peculiarities of the processes and of the analytical work. Apart from in the process itself, it is necessary to consider the statistical error of the correlation function caused by the finite length of the observed signal. Knowledge of the correlation time of the stochastic signal makes it possible to estimate the optimal time-difference between two analyses, depending on the variance of the process and the allowed analytical tolerance.

Der Ablauf chemisch-technologischer Produktionsprozesse wird häufig mittels diskontinuierlich durchgeführter Analysen überwacht. Der Zeitabstand ΔT kann prinzipiell — in Analogie zur Prozeßmeßtechnik — mit Hilfe des Abtasttheorems [1—7] bestimmt werden. Bei der praktischen Anwendung bereitet die Beschaffung der notwendigen a-priori-Informationen (z.B. obere Grenzfrequenz) häufig Schwierigkeiten. Die Eigenart stoffwandelnder Prozesse und die Spezifik der analytischen Arbeitsweise bedingen zusätzliche Besonderheiten: Z.B. treten infolge des häufig notwendigen Chargenbetriebes Diskontinuitäten des Prozeßsignals auf. Weiterhin sind Analysen meist kosten- und zeitaufwendiger als die einfache Messung von Prozeßgrößen. Ziel der

Arbeit ist es, praktikable Lösungswege aufzuzeigen für die Beschreibung charakteristischer Größen des Prozeßsignals und für die Bestimmung der Probenfrequenz unter Berücksichtigung dieser Besonderheiten.

GRUNDLAGEN

Die zeitliche Abhängigkeit der Produktgröße x trägt meist stochastischen Charakter [8]. Für eine wirksame Prozeßkontrolle ist zu fordern, daß zeitlich benachbarte Analysenwerte eine noch signifikante Abhängigkeit voneinander besitzen. Diese Forderung kann in allgemeiner Form durch die Autokorrelationsfunktion R_{xx} beschrieben werden [8]. Bei n diskreten Analysenwerten innerhalb des Zeitraumes T im zeitlichen Abstand ΔT erhält man

$$R_{xx}(j\Delta T) = (1/n) \sum_{i=1}^{n-j} (x_i\, x_{i+j}) \tag{1}$$

wo i = Laufindex der Analysenwerte ($i = 1, 2, 3 \ldots n$) und j = Laufindex der Korrelationsfunktionswerte ($j = 0, 1, 2 \ldots$).

Dabei wird Stationarität des stochastischen Signals vorausgesetzt, d.h., der lineare Zeitmittelwert $\overline{x(t)}$ und der quadratische Mittelwert $\overline{x^2(t)}$ sind im Beobachtungszeitraum konstant. In der Regel arbeitet man mit zentriertem Signal $\overline{x(t)} = 0$, bei ihm klingt mit wachsendem Abstand $j\Delta T$ zweier betrachteter Analysenwerte die Autokorrelationsfunktion auf Null ab, meist mit exponentiellem Verlauf. Nach Abfall auf den $1/e^2$-ten Wert des Anfangswertes ist die Korrelation zwischen zwei Meßwerten zu vernachlässigen, d.h., signifikante Abhängigkeit ist nicht mehr nachweisbar. Die zugehörige Zeit wird als Korrelationsdauer T_c bezeichnet. Wegen der endlichen Zahl von Messungen n stellt $R_{xx}(j\Delta T)$ einen Schätzwert der wahren Autokorrelationsfunktion $R_{xx}(\tau)$ dar. Die (hier interessierende) untere Vertrauensgrenze ergibt sich aus

$$\Delta_u = R_{xx}(j\Delta T) - \lambda_u R_{xx}(0) \tag{2}$$

λ_u folgt der χ^2-Verteilung (Tabelle 1) mit $f = 2\, T/T_c = n\Delta T/T_c$ Freiheitsgraden [8].

TABELLE 1

Untere Vertrauensgrenze $\lambda_u = f/\chi^2(f, \bar{P})$ von $R_{xx}(0)$ bei einer statistischen Sicherheit von $\bar{P} = 0{,}95$ für einseitige Schätzung

f	λ_u	f	λ_u
30	0,6853	200	0,8547
40	0,7173	300	0,8787
50	0,7407	400	0,8936
60	0,7587	500	0,9040
70	0,7732	750	0,9204
80	0,7852	1000	0,9505
90	0,7954	2000	0,9500
100	0,8042	3000	0,9589

ERMITTLUNG DER KORRELATIONSFUNKTION AUS TECHNISCHEN STOCHASTISCHEN SIGNALEN

Stationäre stochastische Signale

Die Berechnung der Autokorrelationsfunktion aus n Wertepaaren (x, y) erfolgt nach vorliegendem Standardprogramm. Das für die Korrelationsdauer T_c gesuchte Vertrauensintervall bestimmt man am einfachsten graphisch (Abb. 1). Zu jedem Punkt der Autokorrelationsfunktion trägt man die Grenze $\lambda_u R_{xx}(0)$ (Tabelle 1) auf. Der zur Ordinate $1/e^2$ gehörige Abszissenwert dieser Kurve ist die gesuchte untere Vertrauensgrenze.

Größenordnung von ΔT

Zur ausreichend genauen Bestimmung der Korrelationsfunktion fehlt häufig jeder Anhaltspunkt für die Größenordnung von $\Delta T < T_c$. Zur groben Abschätzung führt man eine kleine Serie ($n_0 = 10 \ldots 30$) von Analysen durch im vorläufigen Abstand ΔT_0 und trägt $x(t)$ graphisch auf. ΔT_0 darf als brauchbare Ausgangsgröße zur nachfolgenden Bestimmung von T_c betrachtet werden, wenn zwischen jedem der p Extremwerte im Mittel mindestens $\bar{q} \approx 2 \ldots 3$ Analysenwerte liegen. Im Falle $\bar{q} < 2$ erhält man keine auswertbare Korrelationsfunktion. Man muß dann die Analysenserie wiederholen unter Reduzierung (z.B. Halbierung) von ΔT_0, evtl. unter Erhöhung von n_0 (Abb. 2).

Ist umgekehrt $\bar{q} \gg p$, d.h., die Analysenwerte liegen auf einem monoton fallenden oder steigenden Kurvenzug, so war ΔT_0 unökonomisch klein gewählt, weitere Messungen können mit dem doppelten Wert von ΔT_0 enfolgen.

Signalstückelung

Bei Chargenbetrieb liefern die einzelnen stationären Prozeßabschnitte häufig keine ausreichende Anzahl von Analysenwerten, um die Autokorrelations-

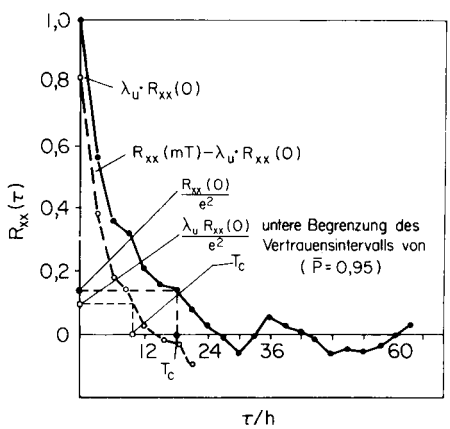


Abb. 1. Graphische Bestimmung der unteren Vertrauensgrenze von T_c : n = 750 Analysenwerte; ΔT = 3h; T_c = 18 h $f = n\Delta T/T_c$ = 125 FG.

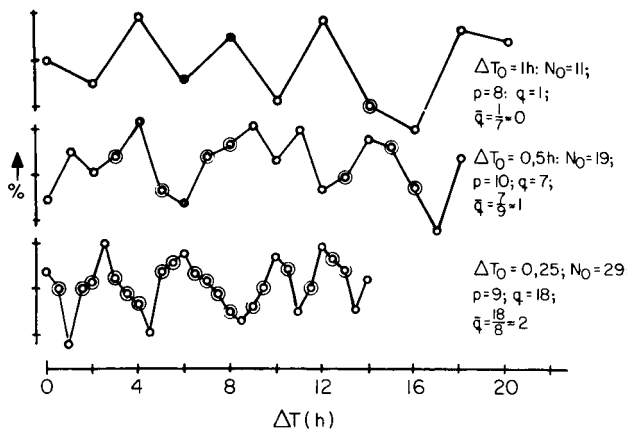


Abb. 2. Graphische Abschätzung von ΔT.

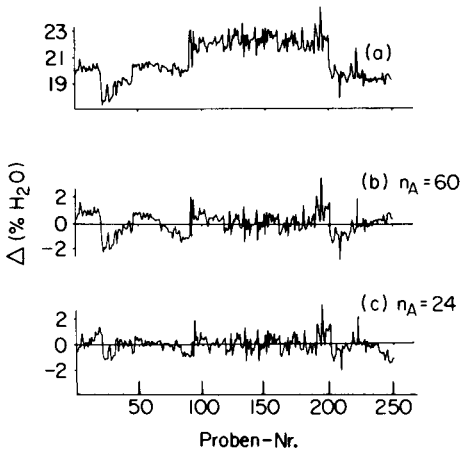
funktion genügend präzise bestimmen zu können. Untersuchungen an simulierten stochastischen Signalen ergaben [9], daß die aus derartigen gestückelten zentrierten Signalen berechneten Autokorrelationsfunktionen nicht signifikant unterscheidbar sind von den Autokorrelationsfunktionen der einzelnen Abschnitte, sofern die Länge jedes Teilstückes T' gegeben ist durch $T' \geqslant 3T_c$. Die für diese Abschätzung notwendige Vorinformation über T_c gewinnt man aus der Korrelationsfunktion des längsten Teilstückes.

Instationäre Prozesse

Technische Prozesse zeigen häufig instationäres Verhalten (z.B. langsam driftenden Mittelwert oder Sprünge des Mittelwertes). Vor Berechnung der Autokorrelationsfunktion muß dann eine Trendkompensation erfolgen, da bei einem driftenden Signal die Autokorrelationsfunktion nicht auf Null abklingt Hierzu läßt sich besonders günstig als Hochpaßfilterung die gleitende Glättung (moving average) [10] einsetzen. Vom jeweils betrachteten Wert $x'_i(t)$ des trendbehafteten Signals subtrahiert man den Mittelwert über n_A Analysenwerte $\bar{x}(n_A)$ im Intervall $i - n_A/2 \ldots i + n_A/2$ und erhält als trendkompensierten Wert

$$x_i(t) = x'_i(t) - [1/(n_A + 1)] \sum_{j=i-n_A/2}^{i+n_A/2} x'_j(t) \qquad (3)$$

Der Mittelungsbereich, d.h. die Wahl von n_A, bestimmt die Filterwirkung der Trendkompensation. Je größer n_A gewählt wird, desto stärker werden die niederfrequenten Komponenten (langsame Trends) eliminiert (Abb. 3). Ein zu geringes n_A unterdrückt jedoch auch die zur Ermittlung von T_c benötigten höherfrequenten Komponenten. Die optimale Wahl von n_A erfolgt zweckmäßigerweise iterativ. Man beginnt mit $n_A = n/4$ Meßpunkten und halbiert solange, bis sich erstmalig eine auf Null abklingende Autokorrelationsfunktion ergibt (Abb. 4).



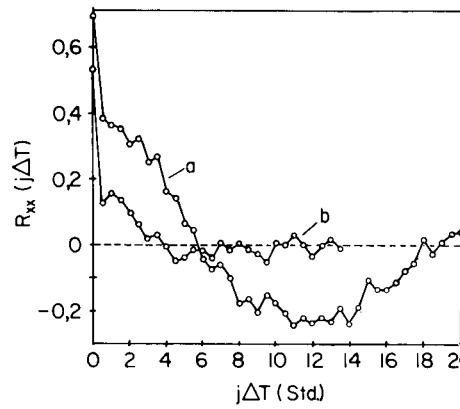


Abb. 3. Verlauf des Prozeßsignals bei iterativer Trendkompensation: (a) vor der Trendkompensation; (b, c) nach der Trendkompensation in Abhängigkeit vom Mittelungsintervall.

Abb. 4. Verlauf der Autokorrelationsfunktion: (a) ohne Trendkompensation; (b) nach Trendkompensation ($n_A = 24$).

Für $i < 1 + n_A/2$ (d.h. am Beginn des Prozeßsignals) bildet man zur Vermeidung von Datenverlusten den zu subtrahierenden Mittelwert nach

$$\bar{x}(n_A) = [1/(i + n_A/2)] \sum_{j=1}^{i+n_A/2} x_j'(t) \tag{4}$$

Zu Beginn und zu Ende der Trendkompensation werden hierdurch die hohen Frequenzen etwas stärker unterdrückt, außerdem werden in diesem Bereich beim trendkompensierten Signal systematische Abweichungen für den Verlauf $x(t)$ verursacht. Beides ist jedoch für die Ermittlung der Korrelationsdauer technologischer Signale vernachlässigbar.

DER OPTIMALE PROBENABSTAND

Für den zeitlichen Abstand zweier Analysen gilt $\Delta T < T_c$. Einen optimalen Probenabstand kann man danach auswählen, daß der Verlauf der Prozeßgröße $x(t)$ zwischen zwei Analysenwerten eine vorgegebene Schranke nicht überschreiten darf. Dazu ist zu untersuchen, welche Abweichungen vom Prozeßverlauf zwischen zwei Probenahmen im Abstand ΔT möglich sind.

Der Verlauf der Prozeßgröße $x(t)$ im Bereich ΔT wird durch zwei diskrete Messungen zur Zeit $i\Delta T$ und $(i + 1)\Delta T$ ermittelt. Die größten Abweichungen sind am wahrscheinlichsten in der Mitte dieses Bereiches, d.h., bei $i\Delta T + \Delta T/2$.

Nach Box und Jenkins [10] erhält man für die Extrapolation der Funktion $x(t)$ von der Stelle $i\Delta T$ auf $i\Delta T + \Delta T/2$ als Erwartungswert der Funktion an der Stelle $(i\Delta T + \Delta T/2)$

$$x(i\Delta T + \Delta T/2) = ax(i\Delta T) \tag{5}$$

mit $a = R_{xx}(\Delta T/2)/R_{xx}(0)$. Unter der Annahme eines exponentiellen Verlaufs von R_{xx} wird damit

$$a = \exp(-\Delta T/T_c) \tag{6}$$

Für die Streuung des Funktionswertes $x(t)$ um diesen Erwartungswert ergibt sich

$$\sigma^2_{\Delta T/2} = (1 - a^2)\sigma_x^2 \tag{7}$$

(σ_x^2 = Varianz des Prozeßsignals $x(t)$). Für den Fall der Interpolation — $x[i\Delta T]$, $x[(i+1)\Delta T]$ bekannt — läßt sich unmittelbar zeigen, daß $\sigma^2_{\Delta T/2} \leqslant (1 - a^2)\sigma_x^2$. Zwischen zwei Probenahmen $[i\Delta T$ oder $(i+1)\Delta T]$ kann die Prozeßgröße vom Mittelwert $\overline{x(t)} = 0$ maximal abweichen um

$$\delta x = |ax(i\Delta T)| + 2\sigma_{\Delta T/2} \tag{8}$$

Diese maximal mögliche Abweichung hängt somit ab vom Verhältnis $\Delta T/T_c$ (Gl. 6) und vom Funktionswert des Prozeßsignals $x[i\Delta T]$ (Abb. 5a). Beispielsweise kann sich bei $\Delta T = T_c$ ein Funktionswert $x[i\Delta T] = \overline{x(t)} = 0$ bis zum Wert $x[t = [(i + 1/2)\Delta T] \approx 2\,\sigma_x$ ändern. Mit abnehmendem Berhältani $\Delta T/T_c$ wird δx immer weniger abhängig vom speziellen Funktionswert $x[i\Delta T]$ (Abb. 5b). Unter Vernachlässigung dieser Abhängigkeit kann man das Vertrauensintervall des Zwischenwertes $x[t = (i + 1/2)\Delta T]$ angeben zu

$$\Delta x = \pm 2(1 - a^2)^{\frac{1}{2}}\sigma_x \quad (P = 0{,}95) \tag{9}$$

Gleichung (9) gilt streng nur für $\Delta T/T_c < 0{,}2$, sie läßt sich jedoch auch für $\Delta T/T_c > 0{,}2$ als Abschätzung nach der sicheren Seite benutzen. Ist bekannt, welche Abweichung Δ zwischen zwei Analysenwerten $x[i\Delta T]$ und $x[(i+1)\Delta T]$ als bedeutungslos angesehen werden darf, so folgt für $\Delta < 2\,\sigma_x$ aus Gl. (9)

$$\Delta T/T_c \approx -\lg[1 - (\Delta/2\,\sigma_x)^2] \tag{10}$$

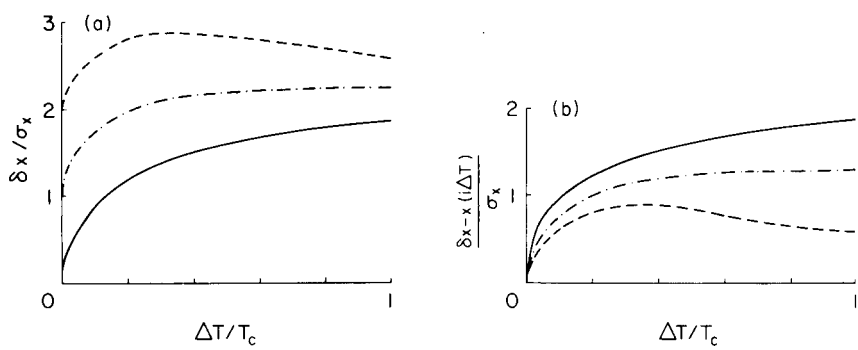


Abb. 5. (a) Prozeßsignal zur Zeit $t = (i + 1/2)\Delta T$ in Abhängigkeit vom Funktionswert und vom Verhältnis $\Delta T/T_c$. (b) Reduziertes Prozeßsignal $x[(i + 1/2)\Delta T] - x[i\Delta T]$ in Abhängigkeit vom Funktionswert. (—) $x[i\Delta T] = 0$; (—·—) $x[i\Delta T] = \sigma_x$; $x[i\Delta T] = 2\sigma_x$.

Der zeitliche Abstand zwischen zwei Analysen hängt ab von der zulässigen Abweichung Δ zweier Analysenwerte und der Streuung der Prozeßgröße σ_x. Die Analyse muß umso häufiger erfolgen, je niedriger das Verhältnis Δ/σ_x liegt, je höher die Gefahr ist, daß der Prozeß eine (z.B. durch Qualitätsnorm festgelegte) kritische Grenze überschreitet. Jedoch soll auch bei Prozessen, die von solchen Grenzen genügend weit entfernt verlaufen, stets $\Delta T \leqslant T_c$ sein.

Geht man bei der Ermittlung des optimalen ΔT von den Kosten aus, die einerseits durch die Überschreitung des Grenzwertes und anderseits durch die Aufwendungen für die Probenahme und -auswertung entstehen, so führt dies zu Resultaten, wie sie in [11] beschrieben wurden.

LITERATUR

1 C. E. Shannon, Communication in the Presence of Noise, Proc. IRE, 37 (1949) 10.
2 A. J. Monroe, Digital Processes for Sampled Data Systems, Wiley, New York, 1962.
3 J. I. Churkin, C. P. Jakolew und G. Wunsch, Theorie und Anwendung der Signalabtastung, VEB Verlag Technik, Berlin, 1966.
4 L. D. Enochson und R. K. Otnes, Programming and Analysis for Digital Time Series Data, Navy Publication and Printing Service Office, Washington, D.C., 1968.
5 G. Wunsch, Systemanalyse, VEB Verlag Technik, Berlin, 1969.
6 J. S. Bendat und A. G. Piersol, Random Data: Analysis and Measurement Procedures, Wiley—Interscience, New York, 1971.
7 M. Schwartz und L. Shaw, Signal Processing, McGraw-Hill, New York, 1975.
8 W. Giloi, Simulation und Analyse stochastischer Vorgänge, Oldenbourg Verlag, München, 1967.
9 U. Vorkel, Berechnung stochastischer Kennfunktionen aus unabhängigen Teilrealisierungen eines stochastischen Signals, Diplomarbeit TH Leuna-Merseburg, Sektion Verfahrenstechnik, 1975.
10 G. Box und G. Jenkins, Time Series Analysis, Holden-Day, San Francisco, 1970.
11 P. J. W. Müskens, Anal. Chim. Acta, 103 (1978) 445.

Now there is a journal for your speculative paper

SPECULATIONS IN SCIENCE AND TECHNOLOGY

If you have a speculative paper, send it to *SPECULATIONS IN SCIENCE AND TECHNOLOGY*, the first international journal devoted solely to speculative papers in the physical, mathematical, biological, medical and engineering sciences.

SPECULATIONS IN SCIENCE AND TECHNOLOGY was launched in 1978 to provide all scientists with a forum for their speculative papers. The editor of the journal defines a "speculation" as an idea which may not be supported by a currently accepted body of experimental or theoretical work. It may, however, subsequently find such support or may be disproven. Its usefulness is that it may help to uncover ideas and procedures of ultimate utility and, in its discussion, clarify current understanding of an idea.

Submissions should be sent in duplicate, together with an abstract of no longer than 150 words, to:

Dr. William M. Honig, Editor
Speculations in Science and Technology
Western Australian Institute of Technology
Perth, S. Bentley 6102
Western Australia

Again, submissions need not be directly based on established bodies of theoretical and experimental work. The efficiency of a paper in explaining an idea or its heuristic value will be the major criterion for acceptance. (Note: No papers on extra-sensory perception or unidentified flying objects will be accepted.) Detailed instructions to authors may be obtained from the publishers or the editor.

SUBSCRIPTION INFORMATION

Volume 2, 1979 (5 issues)
Institutional price: Sfrs. 130.00 (*approx. US$ 72.25) including postage.
Personal price: Sfrs. 60.00 (*approx. US$ 33.25) including postage.

**The US$ price is only approximate and is subject to currency fluctuations. The Swiss franc price is definitive.*

SAMPLE COPIES

Requests for review and free sample copies should be sent to the publishers.

ELSEVIER SEQUOIA S.A.
P.O. Box 851
CH-1001 **Lausanne 1**
Switzerland

498 S

VOL. 112, No. 3 ANALYTICA CHIMICA ACTA September 15, 1979

(Computer Techniques and Optimization, Vol. 3, No. 3)

CONTENTS

Qualitative classification of dithiocarbamate compounds from ^{13}C-n.m.r. and i.r. spectroscopic data by pattern recognition techniques
F. W. Pijpers, H. L. M. van Gaal and J. G. M. van der Linden (Nijmegen, The Netherlands) . . 199

An algorithm for ASTM infrared file searches based on intensity data
K. Tanabe, T. Tamura, J. Hiraishi and S. Saëki (Tsukuba, Japan) 211

A combined forward–reverse library search system for the identification of low-resolution mass spectra
J. Kwiatkowski and W. Riepe (Dortmund, E. Germany) 219

Calculation of the information content of retrieval procedures applied to mass spectral data bases
G. van Marlen (Delft, The Netherlands), A. Dijkstra and H. A. van't Klooster (Utrecht, The Netherlands) 233

The DARC pluridata system: the ^{13}C-n.m.r. data bank
J. E. Dubois and J. C. Bonnet (Paris, France) 245

Strategies in molecular spectroscopic analysis with application of queueing theory and digital simulation
B. G. M. Vandeginste (Nijmegen, The Netherlands) 253

Einsatz von Rechenanlagen in der Emissionsspektrometrie Definition der Probleme und die Schwärzungstransformation
M. Matherny (Košice, Czechoslovakia) 277

Characterization of the flow dynamics of an enzyme reaction system
C. L. Pomernacki, T. A. Brubaker, H. R. Brand and J. W. Frazer (Livermore, CA, U.S.A.). . . 287

Multiple-scanning potentiometric stripping analysis
J. Mortensen, E. Ouziel, H. J. Skov and L. Kryger (Aarhus, Denmark) 297

Ermittlung der Analysenfrequenz bei Diskontinuierlichen prozessanalytischen verfahren
K. Doerffel, G. Lorenz (Merseburg, E. Germany) und I. Tagle (Havanna, Cuba). 313

© Elsevier Scientific Publishing Company, 1979.
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher, Elsevier Scientific Publishing Company, P.O. Box 330, 1000 AH Amsterdam, The Netherlands.

Submission of an article for publication implies the transfer of the copyright from the author to the publisher and is also understood to imply that the article is not being considered for publication elsewhere.

Submission to this journal of a paper entails the author's irrevocable and exclusive authorization of the publisher to collect any sums or considerations for copying or reproduction payable by third parties (as mentioned in article 17 paragraph 2 of the Dutch Copyright Act of 1912 and in the Royal Decree of June 20, 1974 (S. 351) pursuant to article 16 b of the Dutch Copyright Act of 1912) and/or to act in or out of court in connection therewith.

Printed in The Netherlands.