ol. 122 No. 2 June 1, 1980 ISSN 0378-4304
mputer Techniques and Optimization, Vol. 4 No. 2)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

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

PUBLICATION SCHEDULE FOR 1980 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	113/1 113/2	114	115	116/1	116/2	117	118/1	118/2	119	120/1	120/2	121
Section on Computer Techniques and Optimization			122/1			122/2			122/3			122/

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 as designated below for rapid and efficient handling:

Papers from the Americas to: Professor Harry L. Pardue, Department of Chemistry, Purdue University, West Lafayette, IN 47090, U.S.A.

Papers from all other countries to: 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.

American authors are recommended to send manuscripts and proofs by INTERNATIONAL AIRMAIL.

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. 111, p. 343 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 10 volumes in 1980. The subscription for 1980 (Vols. 113–122) is Dfl. 1390.00 plus Dfl. 160.00 (postage) (total approx. U.S. $795.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 122) is Dfl. 139.00 plus Dfl. 16.00 (postage) (total approx. U.S. $79.50). Journals are sent automatically by airmail 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. $77.00), plus Dfl. 10.00 (U.S. $5.00) 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.

ENGINEERS MANAGERS SCIENTISTS

George Washington University
Washington, D.C.
School of Engineering and Applied Science
Announces
a 1980 series of

ADVANCED ENGINEERING SEMINARS

to be given in Berlin at the International Congress Center

Selections from the 1980 Series of 75 Seminars

Seminar No. 522 — **May 12-16, November 10-14**
MICROCOMPUTERS IN CONTROL SYSTEMS

This seminar is designed to familiarize engineers and managers with the capabilities of microcomputers in control applications to replace discrete digital, analog, and electromechanical control elements. Laboratory sessions will be conducted in which the participants will obtain hands-on experience with a microcomputer in several control applications.

Seminar No. 614 — **May 26-30, November 17-21**
MICROPROCESSORS AND MICROCOMPUTERS

The purpose of this seminar is to enable engineers and others without previous experience in digital systems and microcomputers to manage and conduct design work in microcomputer based systems. Participants should gain an understanding of the basic principles of both microprocessors and microcomputers, including how the various products on the market compare with each other, so that informed decisions can be made when choosing hardware.

Seminar No. 418 — **October 6-10**
COMPUTER SYSTEMS FOR WAREHOUSING

This seminar will present a comprehensive, up-to-date look at how minicomputers can be employed in warehousing and distribution applications. Emphasis will be placed on finding practical solutions to space utilization, reducing operating expenses, and improving productivity.

Various kinds of equipment and their operation will be described in sufficient detail that judgments can be made regarding the selection of the most appropriate system for any organization.

Completion of this seminar should enable participants to communicate effectively with equipment vendors and data processing personnel when implementing a computer-based system in their warehouse operations.

Seminar No. 638 — **October 20-24, 1980**
STRUCTURED ANALYSIS, DESIGN AND TESTING OF COMPUTER SYSTEMS

This seminar emphasizes the structured methods used to develop computer systems that work —accurately, reliably, on time, and within budget. This combination of structured analysis, design and testing does for the analysis and design process what structured programming does for the coding process. The seminar provides a framework for the development of systems designs that exactly fulfill use requirements. Evaluation and refinement techniques for the analysis and design stages of system development will be presented along with the structured methods used for planning, managing, and evaluating the integration and testing of computer applications.

Seminar No. 491 — **October 27-31**
COMPUTER GRAPHIC SYSTEMS: DESIGN AND APPLICATIONS

This seminar will provide background in the design of new applications, from general software implementation strategy to possible equipment acquisition. The emphasis will be on relevant concepts and processes rather than on details of graphics hardware or software.

Seminar No. 466 — **October 27-30**
COMPUTER PERFORMANCE EVALUATION

This seminar will cover how and when to use a wide variety of evaluation tools to measure computer performance. The advantages and disadvantages of various software and hardware monitors will be discussed. Methodologies for summarizing, analyzing, and interpreting the large volume of data produced by computer performance evaluation tools will be covered. Emphasis will be placed on methodologies and techniques that are applicable to a wide range of user environments and vendor equipment.

Please send me the brochures of the
CONTINUING ADVANCED ENGINEERING PROGRAM

Name_____
FIRST MIDDLE LAST

Title_____

Organization_____

Address_____

City and Zone_____

County_____

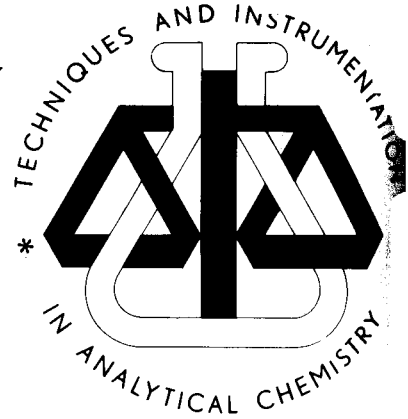
Telephone_____

Please register me for the seminar numbered here. This is tentative. I will, or my office will, confirm at least 2 months in advance. ET

Seminar No._____ Dates_____

ADDRESS TO:
CONTINUING ADVANCED ENGINEERING PROGRAM
George Washington University
Washington, D.C. 20052 U.S.A.
Telephone (202) 676-6106
Telex 64374

Evaluation and Optimization of Laboratory Methods and Analytical Procedures



A Survey of Statistical and Mathematical Techniques

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

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

Techniques and Instrumentation in Analytical Chemistry - Volume 1

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

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

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

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

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

ELSEVIER

P.O. Box 211,
1000 AE Amsterdam
The Netherlands

52 Vanderbilt Ave
New York, N.Y. 10017

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

Analytica Chimica Acta, 122 (1980) 87—94
Computer Techniques and Optimization

A COMPUTER PROGRAM SYSTEM — NEW CHEMICS — FOR STRUCTURE ELUCIDATION OF ORGANIC COMPOUNDS BY SPECTRAL AND OTHER STRUCTURAL INFORMATION

S. SASAKI*, I. FUJIWARA and H. ABE

Toyohashi University of Technology, Tempaku, Toyohashi, Aichi 440 (Japan)

T. YAMASAKI

Mitsui Petrochemical Industry, Co. Ltd., Iwakuni, Yamaguchi 740 (Japan)

(Received 28th September 1979)

SUMMARY

Computer-assisted structure elucidation is improved by the introduction of substructures selected by the user, in addition to analyses of the spectral data of an unknown organic compound. The substructure is called a 'macrocomponent' in the system. The macrocomponent which is input at will is authenticated by comparison with the set of components assembled by the automated analyses of the spectra before it is used for structure construction. It is shown that the introduction of the macrocomponent enhances the correctness and practicality of structure elucidation by computer.

Among several computer-assisted structure elucidation systems [1—3], CHEMICS is one of the most fully automated [4, 5]. In this system, the spectral data (infrared, and proton and carbon-13 n.m.r.) of an unknown organic compound are analyzed automatically and candidate structures for the compound are generated on the basis of the information obtained. The correct answer for the compound is always a member of the group of structures generated. This is the most characteristic feature of the system, because even an operator without experience in organic chemistry can obtain the correct structural formula for an unknown organic compound by introducing appropriate molecular formulas and spectral data. This feature does, however, tend to irritate experienced chemists because of the large number of structures that can be output, many of which are irrelevant to the problem. Chemists almost always have some additional information about unknowns; this may have been obtained by ordinary chemical observation of the sample, by further examination of spectral data, by determination of organic functional groups, etc. Accordingly, CHEMICS has been revised so that it not only accepts the spectral data of an unknown for automated interpretation, but will also accept any size of substructure and/or a skeletal structure available for input from other structural knowledge about the sample. With the addition of this function, CHEMICS provides more accurate structure elucidation, and coincidentally the chemist can participate in the work more closely.

This paper describes the new CHEMICS system which can meet the requirements of both experienced and inexperienced chemists. The new option added for experienced chemists allows the introduction of quite large substructures as additional information at the discretion of the user.

Throughout the system, 189 kinds of specific partial structures, called components, are used (Table 1). The results of the automated spectral data analyses are represented as sets of possible maximum and minimum numbers of the components. Then, all possible sets of components which are consistent with the molecular formula are generated. The candidate structures are finally constructed from those sets.

The basic idea of the new CHEMICS option is that the substructures introduced by users are regarded as if they were additional components; in this sense, such substructures are called macrocomponents. For the input of macrocomponents, a linear notation based on nodes and their connectivity is adopted. This notation is very similar to the natural language for chemists. Users can express any kind of macrocomponent with one or a combination of 36 symbols listed in Table 2. To express a macrocomponent with these symbols, the following rules are applied.

(1) There is no intrinsic limitation on the size of a macrocomponent which can be represented by means of these symbols, but for practical reasons, only macrocomponents which consist of less than 41 symbols can be accepted by the present system.

(2) Symbols which consist of two or more characters can be written in two or more ways, for example, symbol No. 1 can be written as either CH3CH2 or CH2CH3, etc.

(3) Hyphens are used to represent connections of pairs of symbols. A blank space means the termination of any connection.

(4) Modifiers (*n, n = 1, 2, 3, $\cdots$) are used to express cyclic and/or branching structures. Giving the same modifier to a pair of symbols indicates that they are connected to each other. Two or more modifiers can be attached to one symbol for the representation of complicated structures.

(5) A macrocomponent which can be represented by a single symbol should not be represented by two or more symbols. For example, the isopropyl group should be represented only by symbol No. 2 and never by two No. 9 and one No. 11 symbols.

An example of the construction of a macrocomponent for a rather complicated substructure is shown in Table 3. This process is done on paper, and then the resulting macrocomponent code is input via the computer terminal. For simpler macrocomponents, this preliminary paperwork can be omitted after a little experience.

A macrocomponent given for an unknown at the discretion of the user may not always be correct; any user, even a beginner, can input macrocomponents at will and there are few limits on the number, type and complexity of the macrocomponents. Thus the macrocomponent must be authenticated in some way, to avoid abortive computer work. For this purpose, the macro-

TABLE 1

Component table in CHEMICS

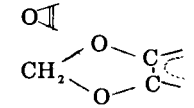
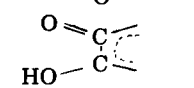
Component	Adjacent group[a]	No.	Component	Adjacent group	No.	Component	Adjacent group[a]	No.	Component	Adjacent group
TERT–BUTYL–	(O)	49	CH3–(CH):	(O)(P)	97	–CH2–	(Y)(T)	145	=C=<ALLENE>	
TERT–BUTYL–	(Y)	50	CH3–(CH):	(A)(A)	98	–CH2–	(K)(K)	146	FURAN(O)	
TERT–BUTYL–	(K)	51	CH3–(CH):	(A)(P)[b]	99	–CH2–	(K)(D)	147	–O–	(K)(K)
TERT–BUTYL–	(D)	52	CH3–(CH):	(Q)(Q)	100	–CH2–	(K)(T)	148	–O–	(K)(O)
TERT–BUTYL–	(T)	53	CH3–(CH):	(Q)(T)	101	–CH2–	(D)(D)	149	–O–	(K)(Y)
TERT–BUTYL–	(C)	54	CH3–(CH):	(T)(T)	102	–CH2–	(D)(T)	150	–O–	(K)(D)
GEM–DIMETHYL–	(O)	55	CH3–(CH):	(C)(O)	103	–CH2–	(T)(T)	151	–O–	(K)(T)
GEM–DIMETHYL–	(Y)	56	CH3–(CH):	(C)(A)	104	–CH2–	(C)(O)	152	–O–	(K)(C)
GEM–DIMETHYL–	(K)	57	CH3–(CH):	(C)(Y)	105	–CH2–	(C)(Y)	153	–O–	
GEM–DIMETHYL–	(D)	58	CH3–(CH):	(C)(K)	106	–CH2–	(C)(K)	154	–CO–	(O)(O)
GEM–DIMETHYL–	(T)	59	CH3–(CH):	(C)(D)	107	–CH2–	(C)(D)	155	–CO–	(O)(Y)
GEM–DIMETHYL–	(C)	60	CH3–(CH):	(C)(T)	108	–CH2–	(C)(T)	156	–CO–	(O)(K)
CH3–(C)–	(O)	61	CH3–(CH):	(C)(C)	109	–CH2–	(C)(C)	157	–CO–	(O)(D)
CH3–(C)–	(Y)	62	:CH.	(O,O,O)	110	CH2=	(C)(C)	158	–CO–	(O)(T)
CH3–(C)–	(K)	63	:CH.	(O,O,A)	111	METHYLENEDIOXY		159	–CO–	(O)(C)
CH3–(C)–	(D)	64	:CH.	(O,O,P)	112	TROPOLONE		160	–CO–	(Y)(Y)
CH3–(C)–	(T)	65	:CH.	(O,A,A)	113	Y–OH		161	–CO–	(K)(Y)
CH3–(C)–	(C)	66	:CH.	(O,A,P)	114	Y–H		162	–CO–	(K)(K)
ISO–PROPYL	(O)	67	:CH.	(O,P,P)	115	CYCLOPROPENONE–H		163	–CO–	(D)(Y)
ISO–PROPYL	(A)[b]	68	:CH.	(A,A,A)	116	T–H		164	–CO–	(D)(K)
ISO–PROPYL	(Y)	69	:CH.	(A,A,P)	117	–CH=<KETENE>		165	–CO–	(D)(D)
ISO–PROPYL	(K)	70	:CH.	(A,P,P)	118	–CH=<OLEFIN>		166	–CO–	(T)(Y)
ISO–PROPYL	(D)	71	:CH.	(Q,Q,P)	119	–OCHO	(O)	167	–CO–	(T)(K)
ISO–PROPYL	(T)	72	:CH.	(Q,T,T)	120	–OCHO	(Y)	168	–CO–	(T)(D)
ISO–PROPYL	(C)	73	:CH.	(T,T,T)	121	–OCHO	(K)	169	–CO–	(T)(T)
CH3O–	(O)	74	:CH.	(C,O,O)	122	–OCHO	(D)	170	–CO–	(C)(Y)
CH3O–	(Y)	75	:CH.	(C,A,O)	123	–OCHO	(T)	171	–CO–	(C)(K)
CH3O–	(K)	76	:CH.	(C,O,P)	124	–OCHO	(C)	172	–CO–	(C)(D)
CH3O–	(D)	77	:CH.	(C,A,A)	125	–OH	(O)	173	–CO–	(C)(T)
CH3O–	(T)	78	:CH.	(C,A,P)	126	–OH	(D)	174	–CO–	(C)(C)
CH3O–	(C)	79	:CH.	(C,Q,Q)	127	–OH	(C)	175	O=C=	
CH3–	(Y)	80	:CH.	(C,Q,T)	128	COOH	(O)	176	T	
CH3–	(D)	81	:CH.	(C,T,T)	129	COOH	(Y)	177	Y	(O)
CH3–	(T)	82	:CH.	(C,C,O)	130	COOH	(K)	178	Y	(Y)
CH3CO–	(O)	83	:CH.	(C,C,A)	131	COOH	(D)	179	Y	(K)
CH3CO–	(Y)	84	:CH.	(C,C,Y)	132	COOH	(T)	180	Y	(D)
CH3CO–	(K)	85	:CH.	(C,C,K)	133	COOH	(C)	181	Y	(T)
CH3CO–	(D)	86	:CH.	(C,C,D)	134	–CHO	(Y)	182	Y	(C)
CH3CO–	(T)	87	:CH.	(C,C,T)	135	–CHO	(K)	183	C	(O)
CH3CO–	(C)	88	:CH.	(C,C,C)	136	–CHO	(D)	184	C	(Y)
CH3CH2–	(O)	89	–CH2–	(O)(O)	137	–CHO	(T)	185	C	(K)
CH3CH2–	(Y)	90	–CH2–	(O)(Y)	138	–CHO	(C)	186	C	(D)
CH3CH2–	(K)	91	–CH2–	(O)(K)	139	–CHO	(CH)	187	C	(T)
CH3CH2–	(D)	92	–CH2–	(O)(D)	140	–CHO	(CH2)	188	C	(C)
CH3CH2–	(T)	93	–CH2–	(O)(T)	141	CYCLOPROPENONE:		189	D[c]	
CH3CH2–	(C)	94	–CH2–	(Y)(Y)	142	:C=<KETENE>				
CH3–(CH):	(O)(O)	95	–CH2–	(Y)(K)	143	:C=<OLEFIN>				
CH3–(CH):	(O)(A)	96	–CH2–	(Y)(D)	144	=C=<KETENE>				

[a]This means the adjacent atom or functional group: saturated oxygen (O); aromatic carbon (Y); carbonyl carbon (K); olefinic carbon (D); acetylenic carbon (T) and saturated carbon (C). [b]A:←O–CO–; P: Y K D T; Q: Y K D. [c]Double bonds.

component is compared in the computer with a set of components resulting from the automated interpretation of the spectral data of an unknown. Macrocomponents that can be shown to be inappropriate, i.e. which conflict with the spectral data interpretation, are rejected by the computer. It is possible to say that automatically selected components are objectively valid on the basis of the concepts of CHEMICS. In the authentication, first the elemental composition of the macrocomponent must be equal to or less than that of the molecular formula of the unknown. Secondly, all the segments comprising the macrocomponent must be present in the set of components assembled by the analyses of spectra of the unknown. Thirdly, the macro-

TABLE 2

Nodes used to express macrocomponents and their codes

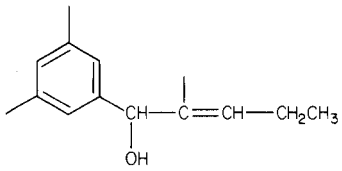
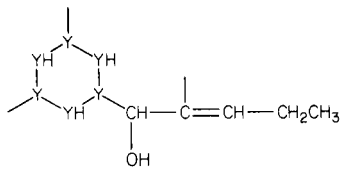
No.	Substructure	Symbol	No.	Substructure	Symbol
1	$CH_3—CH_2—$	CH3CH2	20	$CH{\equiv}C—$	CH#C
2	$\begin{matrix}CH_3\\CH_3\end{matrix}\!\!>CH—$	(CH3)2CH	21	$—C{\equiv}C—$	C#C
3	$CH_3CH<$	CH3CH	22	$—O—$	O
4	$CH_3—\overset{\displaystyle CH_3}{\underset{\displaystyle CH_3}{\overset{|}{\underset{|}{C}}}}—$	(CH3)3C	23		FO
5	$CH_3—\overset{\displaystyle CH_3}{\overset{|}{\underset{|}{C}}}—CH_3$	(CH3)2C	24	$—OH$	OH
6	$CH_3—\overset{|}{\underset{|}{C}}—$	CH3C	25	$—CO—$	CO
7	$CH_3O—$	CH3O	26	$—CHO$	CHO
8	$CH_3CO—$	CH3CO	27	$—COOH$	COOH
9	$CH_3—$[a]	CH3U	28	$—OCHO$	OCHO
10	$—CH_2—$	CH2	29		YH
11	$—\overset{|}{\underset{|}{CH}}—$	CH	30		YOH
12	$—\overset{|}{\underset{|}{C}}—$	C	31		Y
13	$CH_2=$	CH2=	32	$=C=O$	=C=O
14	$—CH=$	CH=	33		CPH
15	$—CH=$[b]	KCH=	34		CPN
16	$>C=$	C=	35		MDO
17	$>C=$[b]	KC=	36		TPL
18	$=C=$	=C=			
19	$=C=$[b]	=KC=			

[a]Methyl connected with unsaturated carbon.
[b]Substructure for ketene structure.

component must be consistent with the members of a molecular subset. Here, a molecular subset is defined as follows: for the plausible components obtained by automated spectral analysis, the consistency of the molecular formula with how many of which components is examined; this set of components, which is equal to the elemental composition of the molecular formula, is called the 'molecular subset'.

TABLE 3

Procedure for constructing a macrocomponent for a substructure

Step no.	Description
	Substructure under consideration: 
1	Selection of appropriate symbols: 
2	Disposition of the symbols in a line in an arbitrary order: Y YH Y YH Y YH CH OH C= =CH CH2CH3
3	Addition of any necessary modifiers to the symbols: Y*1 YH Y*2 YH Y YH*1 CH*2*3 OH C=*3 =CH CH2CH3
4	Connection of symbols by hyphens and removal of excess double bonds: Y*1—YH—Y*2—YH—Y—YH*1 CH*2*3—OH C*3=CH—CH2CH3

RESULTS

The procedure can be explained by an example. For an unknown, CHEMICS assembles 29 components (Nos. 10, 11, 12, 14, 17, 33, 38, 40, 106, 107, 108, 109, 118, 143, 144, 145, 146, 153, 172, 173, 174, 175, 177, 182, 184, 185, 186, 187 and 188; see Table 1) by spectral data analyses of a sample compound with molecular formula $C_9H_{14}O$, and the user wishes to input three macrocomponents, $CH_3—\dot{C}=CH—$, $—\dot{C}=C=CH—$ and $O=C=CH—$, as additional structural information. First, the question of their appropriateness is examined as shown in Table 4. The macrocomponent is degraded into segments which are in the same hierarchy as the components shown in Table 1. Namely, O=C=CH— is degraded to —CH=(ketenic) and O=C=, which correspond to No. 117 and No. 175, respectively. Symbol No. 117 is not present in the set of components assembled by the automated interpretation of spectra, and so this macrocomponent is discarded. The second component, $—\dot{C}=C=CH—$, is degraded to —CH= (No. 118), $—\dot{C}=$ (No. 143) and =C= (No. 144); all of them are present in the 29 components. Then the next check is

carried out: in this case, 5 molecular subsets are possible on the basis of the 29 components and the molecular formula $C_9H_{14}O$ as shown in Fig. 1. It is obvious that construction of an allenic form is impossible by combination of the components in any of the molecular subsets 1—5. Therefore this second macrocomponent is also rejected. The allyl macrocomponent, $CH_3—\overset{|}{C}=CH—$, however, passes through these checkpoints and can be used as an effective tool to avoid construction of inappropriate molecular structure.

Actually, 12 structures are generated by CHEMICS; 3, 1, 2, 3 and 3 from the molecular subsets 1, 2, 3, 4 and 5, respectively (Fig. 2). On introducing $CH_3—\overset{|}{C}=CH—$ only four structures are generated: 0, 0, 0, 2 and 2 from the subsets 1, 2, 3, 4 and 5, respectively. No molecular structure with $CH_3—\overset{|}{C}=CH—$ is constructed by combination of the components in the molecular subsets 1—3. In molecular subset 5, components 143, 118 and 33 are eliminated by the input of $CH_3—\overset{|}{C}=CH—$, therefore the computer does not have to consider all the combinations of the seven components but only the combinations of $CH_3—\overset{|}{C}=CH—$ and the remaining four components Nos. 172, 106, 107 and 12. Thus the five-membered ring is eliminated and the two six-membered rings survive (Fig. 3). Similarly, two six-membered ring structures are generated from the molecular subset 4.

The results for several compounds computed by CHEMICS and CHEMICS with the aid of a macrocomponent are shown in Table 5.

TABLE 4

Examination of macrocomponent by comparison with the components output from processed spectral data. (Numerals 1 and 0 indicate the presence and absence of the said component, respectively.)

No.	Component	Results of automated interpretation of spectra	Macrocomponents user wishes to input		
			$CH_3—\overset{|}{C}=CH—$	$—\overset{|}{C}=C=CH—$	$O=C=CH—$
33	CH_3(D)	1	1	0	0
117	—CH=	0	0	0	1
118	—CH=	1	1	1	0
143	—C=	1	1	1	0
144	=C=	1	0	1	0
145	=C=	1	0	0	0
175	O=C=	1	0	0	1
Another 22 components omitted to save space			Included in 5 molecular subsets	Not included in 5 molecular subsets	False
			True	False	

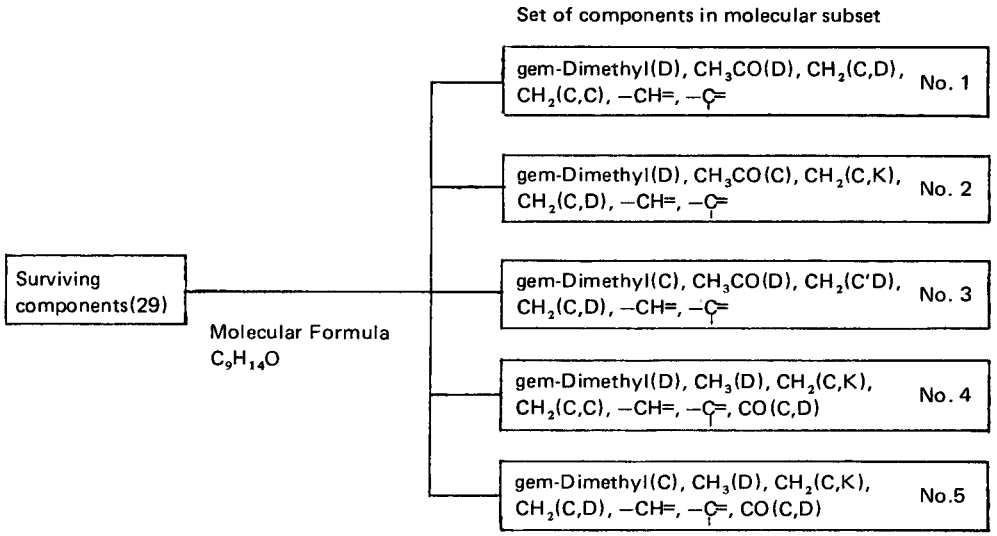


Fig. 1. Molecular subsets for 29 components and $CH_9H_{14}O$. (Sum total of components in each subset is always equal to the molecular formula.)

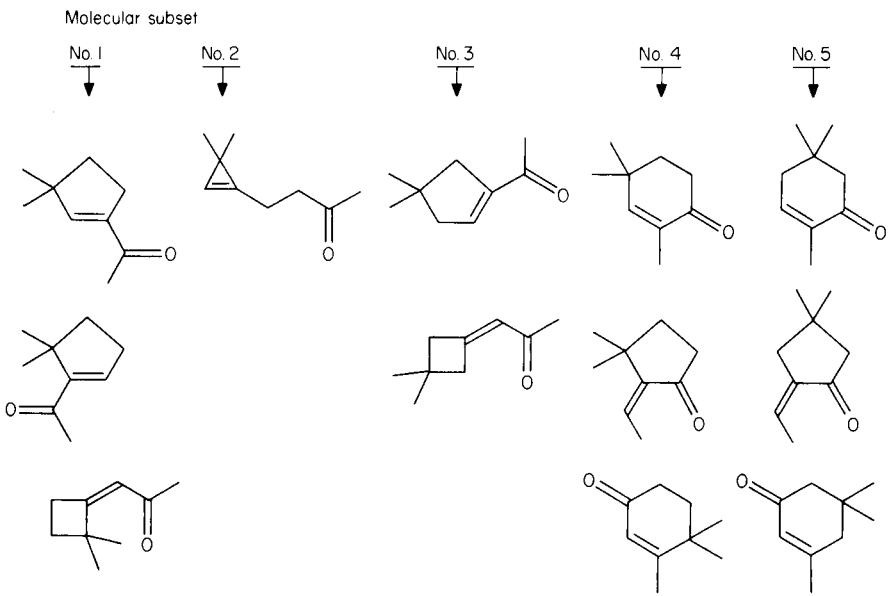


Fig. 2. Structures given for a compound with molecular formula of $C_9H_{14}O$.

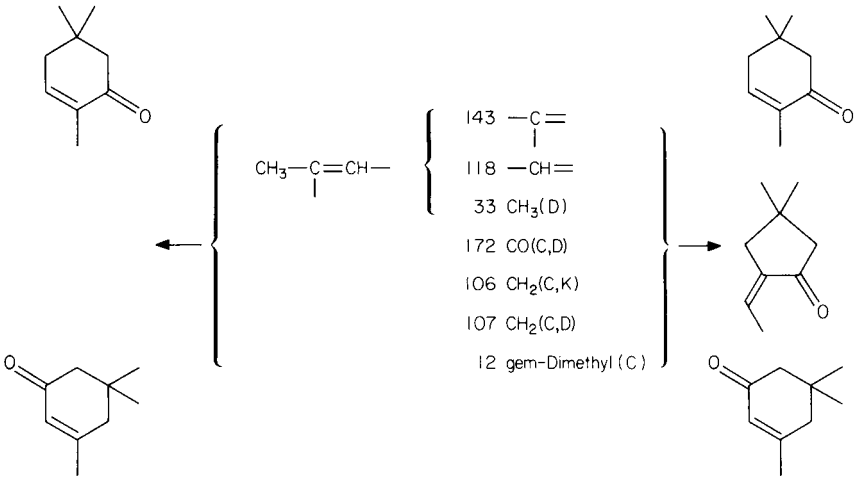


Fig. 3. Elimination of five-membered cyclic structure by introducing a macrocomponent (molecular subset No. 5 in Fig. 1).

TABLE 5

Results for several compounds computed by CHEMICS and CHEMICS with the introduction of macrocomponent. The macrocomponent used is indicated in parenthesis

Compound	Molecular formula	Number of structures	
		CHEMICS	CHEMICS with macrocomponent
m-Xylene	C_8H_{10}	21	4 (>C—C<)
Ethylbenzene	C_8H_{10}	5	1 ($CH_3CH_2C\equiv$)
Coumarin	$C_9H_6O_2$	116	42 (—CH=CH—CO—)
Isophorone	$C_9H_{14}O$	12	4 ($CH_3—\overset{|}{C}=CH—$)
Dicyclopentadiene	$C_{10}H_{12}$	41	20 ($—\overset{|}{C}H—CH=CH—\overset{|}{C}H—$)
Carvone	$C_{10}H_{14}O$	69	6 ($—CO—C(CH_3)=CH—$)
Camphor	$C_{10}H_{16}O$	75	32 ($—C(CH_3)_2—\overset{|}{C}(CH_3)—CO—$)
β-Ionone	$C_{13}H_{20}O$	481	17 ($CH_3—CO—CH=CH—$)

This work was financially supported by the Japanese Ministry of Education, Science and Culture, Grant-in-aid for Developmental Scientific Research No. 384028.

REFERENCES

1 R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, J. Am. Chem. Soc., 97 (1975) 5755.
2 C. A. Shelley, T. R. Hays, M. E. Munk and R. V. Roman, Anal. Chim. Acta, 103 (1978) 121.
3 L. A. Gribov, M. E. Elyashberg and M. M. Raikhshtat, J. Mol. Struct., 53 (1979) 81.
4 T. Yamasaki, H. Abe, Y. Kudo and S. Sasaki, CHEMICS: A Computer program system for structure elucidation of organic compounds, in D. H. Smith (Ed.), Computer-assisted Structure Elucidation, ACS Symposium Series, No. 54, 1977 p. 108.
5 S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito and T. Yamasaki, J. Chem. Inf. Comput. Sci., 18 (1978) 211.

Analytica Chimica Acta, 122 (1980) 95—102
Computer Techniques and Optimization

CHEMICS—UBE, A MODIFIED SYSTEM OF CHEMICS

T. OSHIMA, Y. ISHIDA and K. SAITO

Central Research Laboratory, Ube Industries Ltd., Ube, Yamaguchi 755 (Japan)

S. SASAKI*

Department of Materials Science, Toyohashi University of Technology, Toyohashi, Aichi 440 (Japan)

(Received 28th September 1979)

SUMMARY

Two kinds of contraint have been added to the CHEMICS program system for structure elucidation of organic compounds. One is a limitation on the construction of small ring structures; the other is the introduction of a facility for input of appropriate information at the discretion of the user during the automated analyses of spectral data of unknown compounds. The computation processes for a couple of compounds by the new system are described.

As already described [1, 2] CHEMICS has been endowed with an interactive function by introducing the concept of macrocomponents. Thus not only components given by the automated analyses of spectral data but also macrocomponents input by the user are available for building up candidate structures, and the results of computer-assisted structure elucidation can be dramatically improved. In this paper, another modification of CHEMICS is described. This modified system is temporarily called CHEMICS—UBE. In the NEW CHEMICS system [2] the molecular formula and i.r., H-n.m.r. and ^{13}C-n.m.r. data for an unknown compound are successively compared with the 189 components [2] previously stored in the computer. Lists of surviving components (I, II and III in Fig. 1) are recorded in the computer immediately after the analyses of molecular formula and the three kinds of spectral data. The members in each list of components are the components that are automatically selected given the criterion that all components compatible with the input structural information must be adopted in constructing candidate structures. Rather large numbers of inappropriate components inevitably accompany the appropriate ones. To eliminate unsuitable components, an optional program for confirming the presence or absence of any component in the list at the user's discretion has been added to CHEMICS system. For instance, for tropolone, methylenedioxy and some other unsaturated components (multiple-bond group) can be sought in the list of components (I) resulting from the molecular formula analysis. For example, if the component methylenedioxy is involved in the list, and the user decides that this

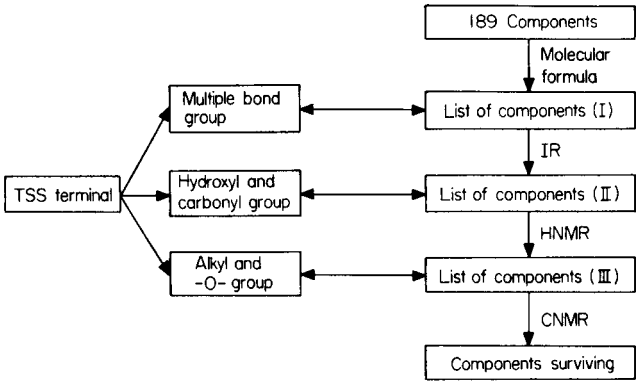


Fig. 1. Mode of interaction in CHEMICS—UBE.

component cannot be present, he can type in NO on the Time Sharing System (TSS) terminal. If the user decides that the component may be present, he types YES. If no definite decision can be made, UNKNOWN is typed in. Even when a component is indicated as present by YES, if this component is rejected in the following analytical step, it will be eliminated from consideration. Components with a hydroxyl and/or carbonyl group and components with alkyl or ethereal oxygen, etc. (alkyl and —O— group) are provided for answer and question for lists II and III, respectively (Fig. 1 and Table 1). The components contained in each group are listed in Table 1.

That all the possible structures consistent with the given structural information are always generated for an unknown is a characteristic feature of the CHEMICS system. Thus, a large number of structures may sometimes be

TABLE 1

Components involved in the three kinds of group in Fig. 1

Group	Component		
Multiple bond	Tropolone (112) Methylenedioxy (111) Cyclopropenone (115, 141) Furan (146)	Ketene (117, 143, 144, 175) Allene (142, 145) Aromatic ring (114, 172—182)	C=C double bond (110, 118, 143, 189) Triple bond (116, 176) Ring structure (input max. and min. of ring member)
Hydroxyl and carbonyl	Phenolic OH (113) Alcoholic OH (126, 127) Acetyl (35—40)	Acetoxyl (35) Formate (119—124) Carboxyl (128—133)	Formyl (134—140) Ester or Lactone (147—152, 154—159)
Alkyl and —O—	t-Butyl (1—6) iso-Propyl (19—25) gem-Dimethyl (7—12) Ethyl (41—46)	Methoxyl (26—31) R—O—O—R′ (29, 119, 125, 148) R—O—CH_2—OR′ (89)	R—CH(OR′)—OR″ (47, 48, 50, 64, 66, 74, 75, 76, 77) R—O—CH(OR′)—OR″ (62, 63, 65, 68)

Numeral in parentheses indicates code number of component (see Component Table in the preceding paper).

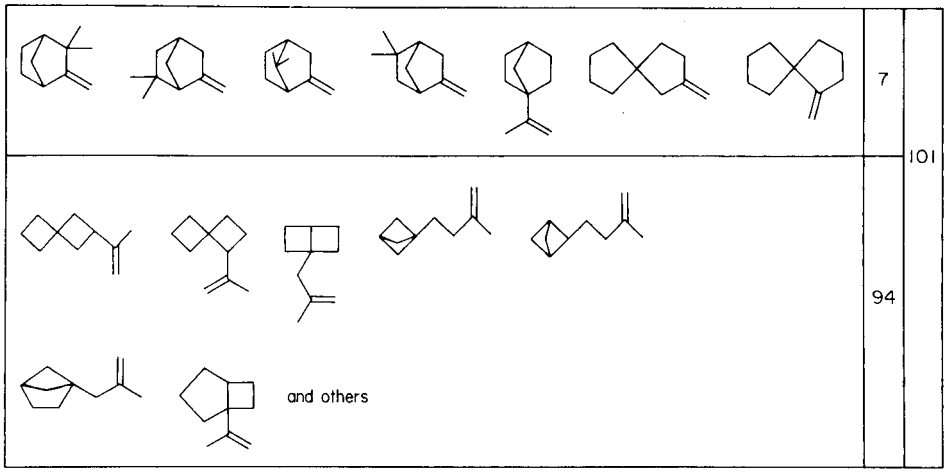


Fig. 2. Structures (101) given for the data of camphene by CHEMICS.

generated for some compounds, even if structural information is available from i.r., H-n.m.r. and ^{13}C-n.m.r. In the case of camphene ($C_{10}H_{16}$), for example, 101 cyclic formulas are constructed based on the data from i.r., ^{13}C-n.m.r. and H-n.m.r., as shown in Fig. 2. Of these 101 structures, only seven formulas contain a five or six-membered ring, and the others are so-called small ring structures. It is not always necessary, in practice, to take account of such small ring structures. Thus an additional program with a function for suppressing the construction of small structures was connected to the structure-building program in CHEMICS. Information on the ring size can be input to the computer through the TSS terminal as part of the chemical evidence that can be input at the discretion of the user. A couple of parameters, the maximum and minimum number of ring members, are used to express the decision of the user. In order to prohibit construction of cyclic structures containing four members or less, the minimum value of parameter defines five. Inputting UNKNOWN also indicates that the small rings should be excluded. The epoxide and γ-lactone groups are not regarded as small ring structures for such purposes.

EXAMPLES OF COMPUTER-ASSISTED STRUCTURE ELUCIDATION OF UNKNOWN COMPOUNDS BY CHEMICS—UBE

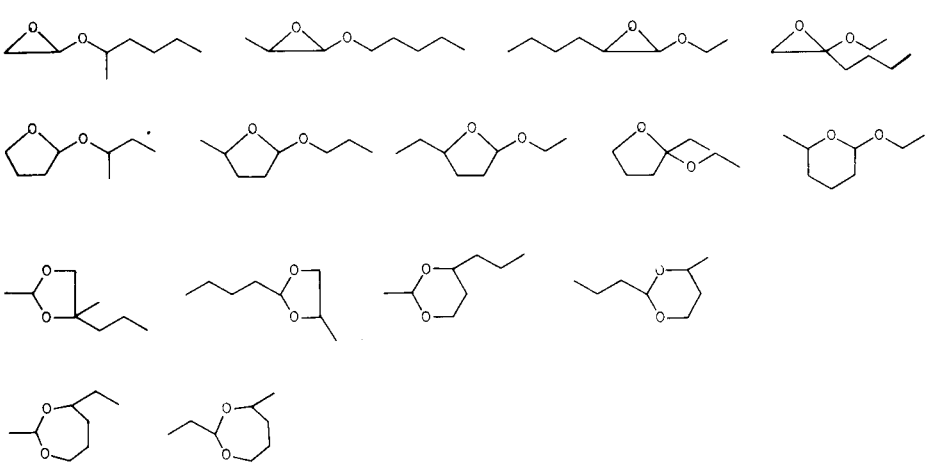
The computational process for a dimer of n-butanol by the system is indicated below. The structural information for this compound is given in Table 2. When the structure elucidation is based on the molecular formula and the three kinds of spectral data, the finally surviving components lead to the generation of 23 and 136 structures for the compound (left side, Fig. 3). In CHEMICS—UBE, two questions regarding the presence of a C=C double bond and the size of the ring structure in the multiple bond group (Table 1),

TABLE 2

Structural information for an n-butanol dimer (molecular formula: $C_8H_{16}O_2$)

H-n.m.r.											
No.	1	2	3	4	5	6	7	8	9	10	11
Ppm	4.60	4.55	4.50	4.19	4.17	4.10	4.04	3.83	3.80	3.75	3.70
Area	155	406	139	98	104	315	273	289	306	394	402
Height	23	37	19	14	14	22	21	22	23	38	31
No.	12	13	14	15	16	17	18	19	20	21	22
Ppm	3.64	1.67	1.61	1.56	1.50	1.45	1.24	1.19	0.98	0.90	0.84
Area	112	285	789	1532	1388	520	987	912	788	1110	500
Height	15	28	45	71	68	55	149	147	70	102	28
I.r.											
No.	1	2	3	4	5	6	7	8	9	10	11
Cm^{-1}	2955	2875	2850	1465	1385	1325	1165	1135	1120	1110	985
Intensity	69	49	54	33	54	32	73	74	62	65	52
^{13}C-n.m.r.											
No.	1	2	3	4	5	6	7	8			
Ppm	13.9	17.4	21.7	33.1	37.2	66.5	72.6	101.9			
Intensity	3760.0	5100.0	5813.0	4947.0	3873.0	5109.0	3723.0	3832.0			

are referred to 69 components and to the analytical result for the molecular formula; this reduces the number of components to 53. The i.r. analysis diminishes these 53 components to 29; and a question regarding the presence of an alcoholic hydroxyl group then eliminates one component from the 29. The H-n.m.r. analysis reduces 28 to 17 components which are further reduced to 16 by questions regarding the presence of an ethyl group, —O—O— bonding and acetal linkage. Finally, ^{13}C-n.m.r. picks 13 components from these 16 components which have been chosen by the analyses of molecular formula and the two kinds of spectra, with chemical information input by the user as desired. Fifteen structures are built up based on those 13 components (right side, Fig. 3). These structures are:



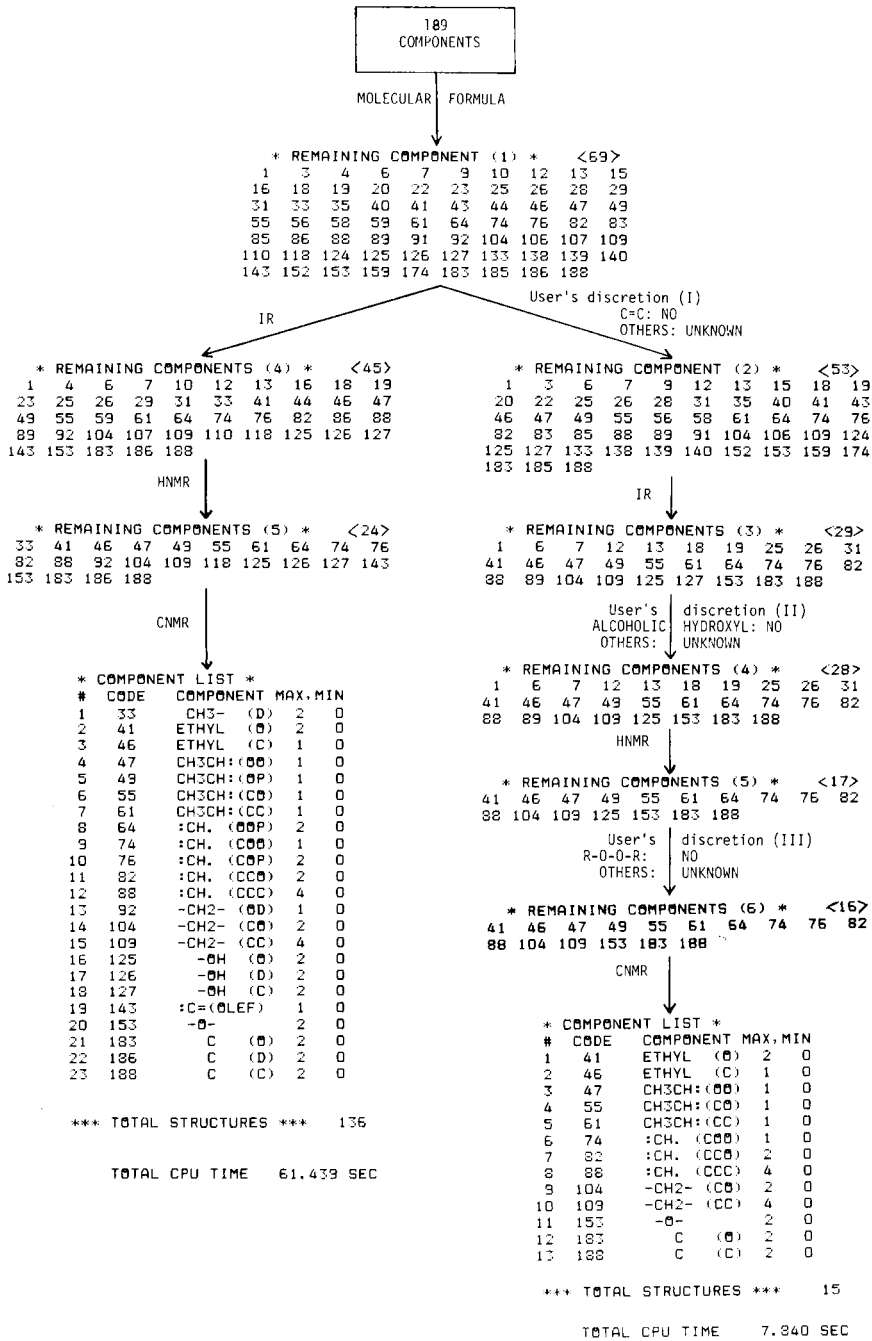


Fig. 3. Computation processes for a dimer of n-butanol by CHEMICS (left) and CHEMICS--UBE (right). Components are expressed by their code numbers (see Component Table in the preceeding paper). Numerals in < > indicate the number of components surviving at each step.

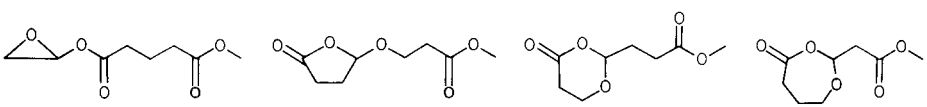
TABLE 3

Structural information for the minor product

H-n.m.r.								
No.	1	2	3	4	5	6	7	8
Ppm	5.61	5.57	5.53	4.08	4.00	3.92	3.69	2.67
Area	180	490	310	620	850	690	3240	955
Height	10	19	13	10	20	15	350	38
No.	9	10	11	12	13	14	15	
Ppm	2.61	2.55	2.47	2.34	2.26	2.24	2.14	
Area	1090	820	1340	725	480	510	360	
Height	51	27	24	16	14	14	9	
I.r.								
No.	1	2	3	4	5	6	7	8
Cm^{-1}	2950	2845	1780	1735	1430	1350	1270	1150
Intensity	76	55	92	93	76	79	78	89
No.	9	10	11	12	13	14	15	
Cm^{-1}	1110	1030	955	920	860	790	660	
Intensity	83	81	81	70	60	48	53	
C-n.m.r.								
No.	1	2	3	4	5	6	7	8
Ppm	26.8	28.8	34.7	51.8	64.9	104.2	171.5	176.6
Intensity	3750.0	4150.0	4100.0	1720.0	3480.0	2900.0	1410.0	1400.0
Off-resonance data	T	T	T	Q	T	D	S	S

Another example is based on a study of cis-1,4-polybutadiene treated by ozonolysis followed by methylation with diazomethane, which gives dimethyl succinate and an unidentified minor product. The minor compound was separated and purified by g.l.c. and the i.r., H-n.m.r. and ^{13}C-n.m.r. spectra were measured. Mass spectrometric data indicated that the molecular formula was $C_8H_{12}O_5$. The spectral data are given in Table 3.

The total of 189 possible components are diminished to 31 by the analyses of spectral data and by the insertion of information at the discretion of the user as shown in Fig. 4. As a result, only four structures are generated based on the 31 components; one of these, with a γ-lactone structure is proven to be the correct structure by other chemical and physico-chemical observations. The four structures are as follows:



The final outputs in the system are displayed two-dimensionally on the CRT by the program kindly provided by Dr. Carhart [3].

The results for a variety of compounds are summarized in Table 4.

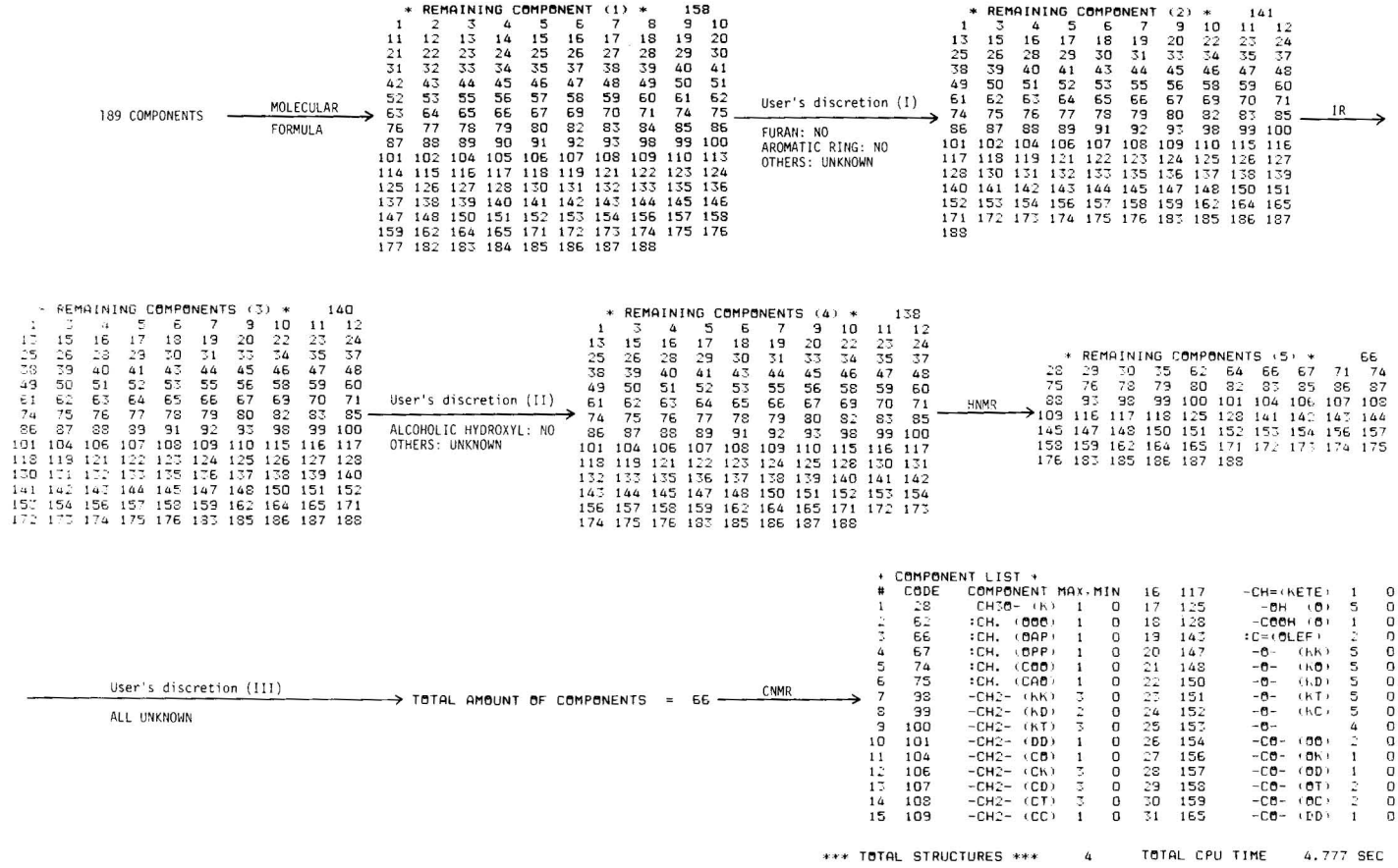


Fig. 4. Computation processes for an ozonolysis product of cis-1,4-polybutadiene CHEMICS—UBE.

TABLE 4

Number of structures (n) generated by CHEMICS—UBE for various organic compounds

Compound input	n	Compound input	n	Compound input	n
(1) Aromatic compounds (Presence of benzene ring or phenolic hydroxyl is suggested by user)					
Phenol	1	Veratrole	3	*o*-n-Propoxyphenol	6
o-Cresol	3	Piperonal	2	Glycol-*o*-toluylether	6
p-Cresol	3	Vanillin	10	*o*-Allyloxytoluene	6
Benzo-1,3-dioxol	1	*o*-Vanillin	10	2-Allyl-6-methylphenol	10
Guaiacol	3	*o*-Allylphenol	3	Safrole	2
2,3-Xylenol	6	*p*-Allylphenol	4	*p*-Eugenol	10
2,4-Xylenol	6	Oxychavicol	6	*o*-Diethoxybenzene	12
Protocatechualdehyde	6	*o*-Allyloxyphenol	6	2,5-Di-tert-butylhydroquinone	11
(2) Alicyclic compounds (Construction of small rings is prohibited by user)					
1,3-Cyclohexadiene	2	Cyclohexylmethanol	2	Cyclohexyl propionate	18
1,4-Cyclohexadiene	1	1,2-Cyclohexane-dicarboxylic		3-Cyclohexylpropionic acid	3
1,2-Cyclohexanedione	3	anhydride-4-ene	3	Cyclohexyl butyrate	12
Cyclohexene	2	1,2-Cyclohexanedicarboxylic		Isopropyl cyclohexane-	
Cyclohexeneoxide	2	anhydride	5	carboxylate	19
Cyclohexanone	1	Methyl cyclohexanecarboxy-		Cyclohexyl isovalerate	17
Cyclohexanol	4	late	2	Cyclohexylbenzene	1
1,2-Cyclohexanediol	6	Cyclohexyl acetate	7	4-Cyclohexylphenol	3
1,3-Cyclohexanediol	9	Cyclohexylacetic acid	3	2-Cyclohexylcyclohexanone	3
Cyclohexene-4-carboxylic		Ethylcyclohexane	2	Diethyl cyclohexane-1,2-	
acid	5	2-Cyclohexylethanol	3	dicarboxylate	3
Cyclohexanecarboxylic acid	4	1,1-Dimethoxycyclohexane	1	4-Cyclohexylcyclohexanol	7
(3) Terpenoids (Suggestions by user are omitted because of the immense variety)					
3-Hexene-1-ol	1	Linalool	9	Pulegone	9
Cyclotene	2	α-Terpineol	5	Fenchone	6
1-Octene-3-ol	1	Citronellal	16	Ascaridol	6
6-Methylhept-5-en-2-one	2	Menthone	14	d-Carvone	9
Citronellol	10	α-Limonene	5	Bornyl acetate	10
l-Menthol	22	β-Pinene	22	Isobornyl acetate	10
Borneol	10	Δ-3-Carene	46	Linalyl acetate	9
1,8-Cineol	2	Camphene	7	Farnesol	20
Geraniol	35	Camphor	5	Perillaldehyde	6

REFERENCES

1 T. Yamasaki, H. Abe, Y. Kudo and S. Sasaki, CHEMICS; A Computer program system for structure elucidation of organic compounds, in D. H. Smith (Ed.), Computer-assisted Structure Elucidation, ACS Symposium Series, No. 54, 1977, p. 108.

2 S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito and T. Yamasaki, J. Chem; Inf. Comput. Sci., 18, (1978) 211. S. Sasaki, I. Fujiwara, H. Abe and T. Yamasaki, Anal. Chim. Acta, 122 (1980) 87.

3 R. E. Carhart, J. Chem. Inf. Comput. Sci., 16 (1976) 82.

Analytica Chimica Acta, 122 (1980) 103—115
Computer Techniques and Optimization

KISIK — A COMBINED CHEMICAL INFORMATION SYSTEM FOR A MINICOMPUTER†

J. ZUPAN*, M. PENCA, M. RAZINGER and B. BARLIČ**
Boris Kidrič Institute of Chemistry, Ljubljana (Yugoslavia)
D. HADŽI
Faculty for Natural Sciences, Edvard Kardelj University, Ljubljana (Yugoslavia)
(Received 7th November 1979)

SUMMARY

A minicomputer-oriented chemical information system (CIS) based on three different spectrometries, i.e. infrared, mass, and ^{13}C-n.m.r., is described and discussed. The system has roughly the same characteristics as CIS's implemented on large mainframe computers: substructure search, library searches on various files, file manipulation, statistical handling of retrieved data, etc. The source package is very suitable and simple for moving the entire CIS from one computer to another. In addition, the system has the option UPDATE that enables the user to create his own files and modify them easily; this is rather difficult and expensive to implement on larger systems because of the very high disk-space price to frequency-of-access ratio. However, the quantity of data is strongly dependent on the disk space. At the moment the system handles 1016 compounds, each of which is described by a chemical name, molecular formula and weight, two-dimensional structure image, infrared, mass, and ^{13}C-n.m.r. spectra. All these data for one compound are linked on-line via the identity number of the compound so there is no delay in accessing any of the items mentioned. The entire data bank together with the program package has a 1.8 Mbyte requirement which fits well within the 2.5 Mbyte space available on the small disk used by a PDP 11/34 minicomputer.

Recently, much has been written on the usefulness, benefits and requirements of chemical information systems (CIS's). Reviews and descriptions of different systems are available in the literature [1–6]. The aim of this paper is not to elaborate on all these features but is directed to the description of solutions implemented on a CIS that was designed specially for minicomputers with particular attention to the economy (of space and time) of the data banks and algorithms employed. The CIS in question, KISIK, has been operating in the analytical laboratory for over a year on the minicomputer PDP 11/34 [7]. The general scheme is shown in Fig. 1. (KISIK is the acronym for the Slovenian words meaning "Chemical Information System of the Institute B. Kidrič". The word KISIK literally means oxygen.)

---

†This paper was presented in part at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.
**Present address: Elektrotehna, TOZD Digital, Ljubljana, Yugoslavia.

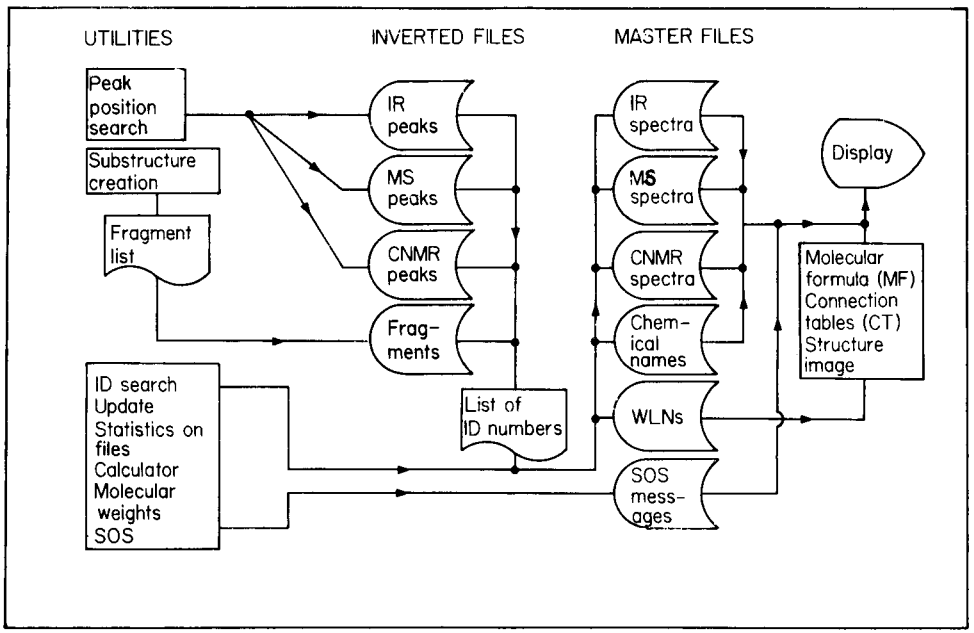


Fig. 1. Organization scheme of the CIS KISIK running on PDP 11/34.

The most important characteristics, including the chemical aspects, that make an information system very useful can be summarized as follows: (a) implementation on a large mainframe computer (with large direct-access storage) accessible via (worldwide) phone networks; (b) attractive choice and content of data collections; (c) freedom of accessibility of data files for different users; (d) restriction of actual updating to the authorized groups responsible for the maintenance of particular data files. However, such characteristics may not be attractive to all users; in fact, many users have very specific requirements incompatible with those mentioned. The problems concerning the organization and use of data files are far greater than the problems of the search system itself. Users, especially industrial ones, are generally very concerned with the security and privacy of data banks even if the quantity of data is small. The commonest conflicting requirements may be listed as follows: (a) the possibility of creating private files with exclusive right of access; (b) updating and maintenance of private files exclusively for their creators; (c) invulnerability to breakdown of public facilities (phone networks, electrical power, computer maintenance, etc.) especially in third-world countries; (d) the possibility of communication in the user's native tongue in interactive work, particularly for less skilled workers; (e) easy transfer to new equipment if the CIS is installed in a stand-alone minicomputer to meet the above requirements. Considering this second group of requirements, and thus trying to satisfy as many potential users as possible and to spread the chemical information flow as widely as

possible, it seems justifiable to attempt to build a powerful version of CIS that will run on a medium-size minicomputer.

The basic flexibility in procedural flow and the resulting file manipulation of large CIS's should obviously not be sacrificed in a minicomputer version. The only feature of the large CIS's which can be curtailed to save space is the number of records in the data banks. Consideration of the number of spectra contained in some published special collections [8—12] and the results of questioning potential users led to the conclusion that the number of spectra in the most frequently used data bases averages about 1 000. This figure, of course, does not imply that larger data files are of no interest to these users. The problems are that large data banks as a whole are less frequently used by specialized workers and that these banks do not normally cover the specific needs of such workers.

The second aspect of space minimization, i.e., the strict compression of all permanently used items, must also be considered. In order to economize the system (not the data banks), the following problems must be overcome: (a) organization of overlay structure at the time of the system generation; (b) compression of internal spectra storage; (c) compression of structure and fragment storage; (d) computation of the structure images of the stored compounds on-line without storage; (e) efficient command language.

The solutions to some of these problems in the interactive, multi-user KISIK running on the minicomputer PDP 11/34 with 48K word memory (28K words for KISIK) are described below.

GENERATION OF THE SYSTEM

The entire system software consists of distinct primary parts, i.e., master data banks, high-level and assembler programs (for data compression, creation of inverted files and the CIS itself), and textual descriptions of procedures and subprocedures needed to set up each part of the entire working system. In order to create the entire system from its elements, the following sequence of logical steps is essential.

(1) Transfer of master data banks and high-level programs on the tapes to card image form.

(2) Writing the corresponding assembler programs for bit manipulation. The precise goal (input and output specification for these routines) is fixed in textual description.

(3) Compilation of software parts for creating compressed and inverted data files.

(4) Creation of compressed and inverted data files. Because both types of file must be available for random access for fast retrieval during interactive sessions, the high-level source programs must be checked to ensure correct connections. Even in high-level programming the descriptions of random-access files are highly machine-dependent.

(5) Adaptation of the high-level CIS software to certain machine-dependent requirements.

(6) Creation of the overlay structure (overlay description) if possible.

(7) Compilation of CIS software in connection with the overlay structure and the creation of the operative version of the CIS.

A very important part of the complete software for any complex program package is a textual description of all these steps. Without such descriptions, it will be very difficult if not impossible, for example, to build an overlay tree structure of subroutines from the list of source programs only or to write the assembler programs for very precisely defined tasks. In the present version of KISIK running on PDP 11/34 under OS RSX11M, the above-mentioned tasks were implemented by using command files which direct the operational system to perform all procedures and subprocedures necessary to generate the entire system. There are only two main command files for KISIK: one contains the system commands for compilation and creation of compressed and inverted files from the original data; the other contains the system commands that lead from the area for the compilation of source programs, creation of library and temporary files and generation of overlay connections, to the KISIK execution.

COMPRESSED STORAGE OF SPECTRA AND STRUCTURAL FRAGMENTS

Master data files that contain numerical information on chemical compounds require much storage space. Each chemical compound described by its chemical name, molecular formula, molecular weight and full two-dimensional structure image, and characterized by three types of spectrum (infrared, mass, and ^{13}C-n.m.r.) requires about 5 Kbytes of storage. Thus for 1000 compounds, about 5 Mbytes are needed which is a fairly large amount of space for a minicomputer. A compressed form for storage is therefore needed for direct access. The compression scheme for all three types of spectrum (i.r., mass, and ^{13}C-n.m.r.) used in KISIK is shown in Fig. 2. Because of their nature (sharp, well-defined peak positions with given relative intensities) the compression of mass and n.m.r. spectra is straightforward. However, experimental i.r. spectra cannot be stored faithfully enough by using only peak positions and intensities. Hence, two additional parameters were employed: the half-width of the peak (HW) and the type of peak shape (Lorentzian or Gaussian). All four parameters (position, intensity, HW and type) are used to calculate the spectrum each time the spectrum has to be displayed on the video terminal. A spectral curve calculated from the compressed form is compared to the real spectrum in Fig. 3. The similarity is only fair, and the user is advised to check the real spectrum in the printed catalog. The catalog number from which the image was derived is displayed along with the spectrum. In spite of some shortcomings, the saving factor of approximately 10 (880 bits compared with about 10 000 bits required by full curve digitization) is impressive, especially if relatively low-grade spectra will suffice, which is normally true for compound identification. There are two advantages in this compression. First, the two most important parameters (position and intensity) of each peak are already known (with some degree of precision)

and no special program is necessary for peak detection, which is particularly cumbersome for infrared spectra. Secondly, the updating or correction of the spectra is fairly easy [13] and can be invoked at any step, even for checking the actual data.

In order to make the substructure search feasible and rapid, two requirements must be met: each compound in the data file must have a well defined structure stored, and there must be an algorithm for decomposing any of those structures to a number of fragments according to which the inverted file entries of structural features can be formed or calculated (hashed).

It can safely be said that the connection table (CT) of Morgan [14] is one of the best two-dimensional descriptions of the topology of molecular structure. Its only possible shortcoming is its rather large space requirement: if the computer space on drums is not the bottleneck of the information system, the CT's can be stored in a format suitable for random access and all the other structural features (molecular formula and weight or structure image) can easily be derived from it. However, if further space-conserving steps are required, the CT's can be derived from some sort of condensed linear notation [15—18]. In the CIS KISIK, a minicomputer version of WLN-to-CT conversion program is applied to generate CT's and ring description tables from Wiswesser line notation (WLN) [15]. Because of the limitation on the complexity of this program [19], compounds having WLN's that cannot be decoded by the program are specially marked (asterisked) to direct the program control to a short oblique file where pre-made CT's for such cases are stored. In the present file of 1016 compounds, only one

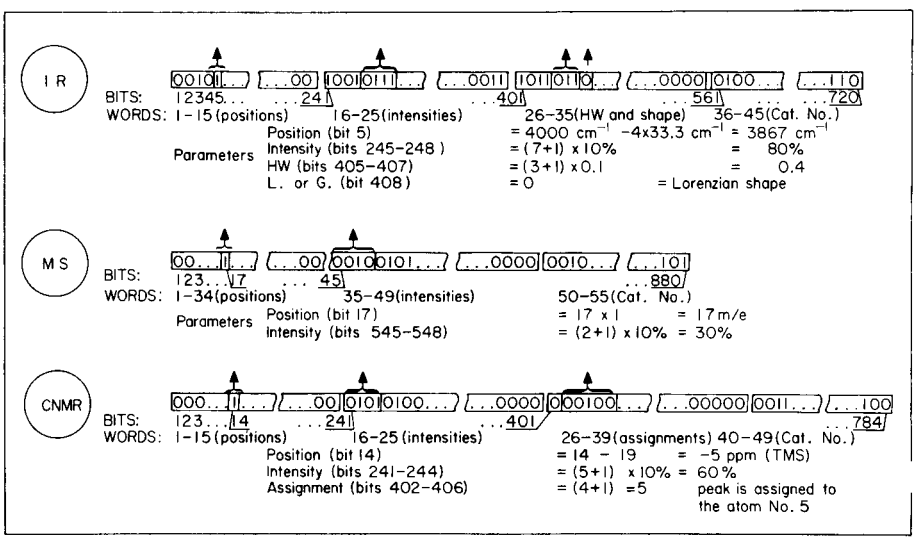


Fig. 2. Compression of infrared, mass and ^{13}C-n.m.r. spectra in the master files from which the spectral images are obtained and displayed on a graphic terminal. If the user does not have the graphic terminal, only the numeric data are displayed.

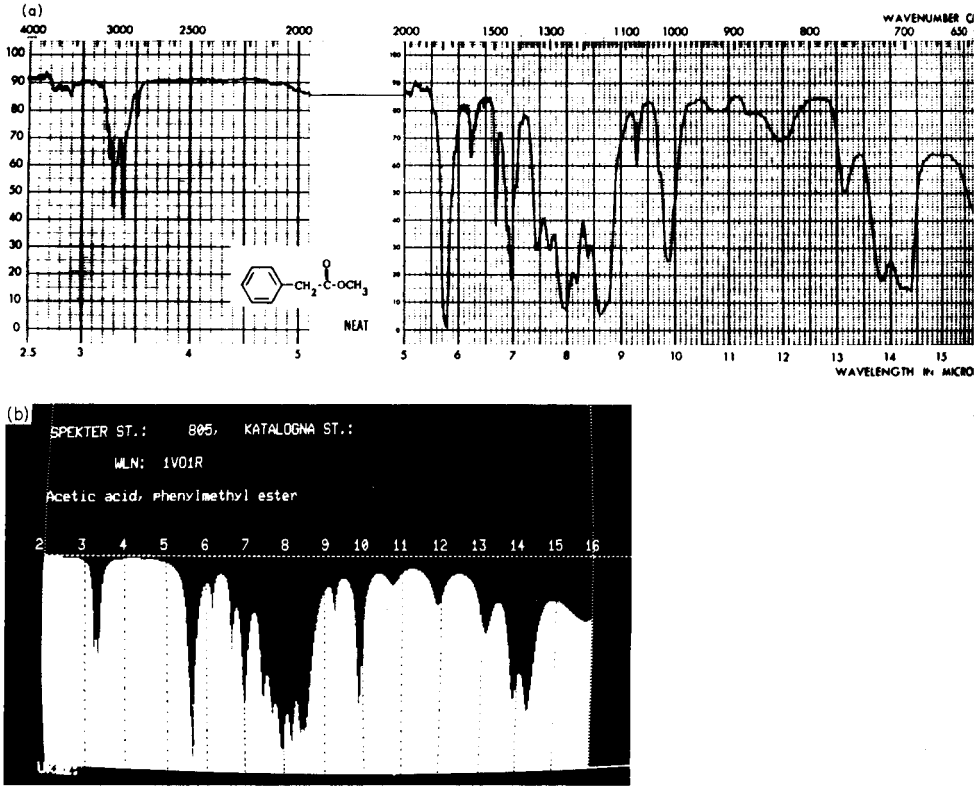


Fig. 3. Comparison between the real i.r. spectrum for phenylmethyl acetate from the catalog (a) and the spectrum calculated from the parameters stored in the compressed form (b) shown in Fig. 2.

compound, adamantane has too complex a WLN to be decoded by the routine mentioned.

The further procedures, i.e., calculation of molecular formula or construction of structural image, are the same regardless of whether the CT was obtained from WLN or from the file, or was generated by the user choosing the option GENESIS.

The second part, decomposition of any structure to its elemental fragments, is relatively simple after an exact definition of what these fragments are has been made. Many different fragment descriptions and definitions are now used in CIS's. In KISIK, the following rules are applied in order to obtain the fragments for any given chemical structure (or substructure) described by the CT:

(a) each fragment is composed of the central atom, its first neighbours and bonds connecting them to the central atom;

(b) each non-hydrogen atom is considered only once as a central atom of a fragment;

(c) eight types of central atom are distinguished

0 carbon	2 oxygen	4 phosphorus	6 any halogen
1 nitrogen	3 sulphur	5 boron	7 all others

(d) two types of neighbour are distinguished

0 carbon	1 all others

(e) eight types of bond between the central atom and its neighbours are coded differently

0 no bond	4 single bond in the ring
1 single bond in the chain	5 double bond in the ring
2 double bond in the chain	6 triple bond in the ring
3 triple bond in the chain	7 ring alternating bond

(f) the type of the central atom is coded first and then the types of bond and neighbour are coded sequentially;

(g) the sequence of neighbours around the central atom is estimated by taking first the highest bond number and if the alternative choice still remains the non-carbon atoms are considered before the carbons.

When these rules are applied, any fragment can be packed into a 16-bit word very efficiently. Figure 4 shows the bit ordering for two fragments. However, some difficulties in coding four or more neighbouring fragments can arise when only the 16-bit strings are used. Fortunately, such fragments are few. In the present scheme, the 16th bit that is normally not used in smaller fragments serves for the assignment that the fragment describes a central atom with more than three neighbours (which are of course described as explained).

The advantages of this coding of fragments are manifold. Thus, the

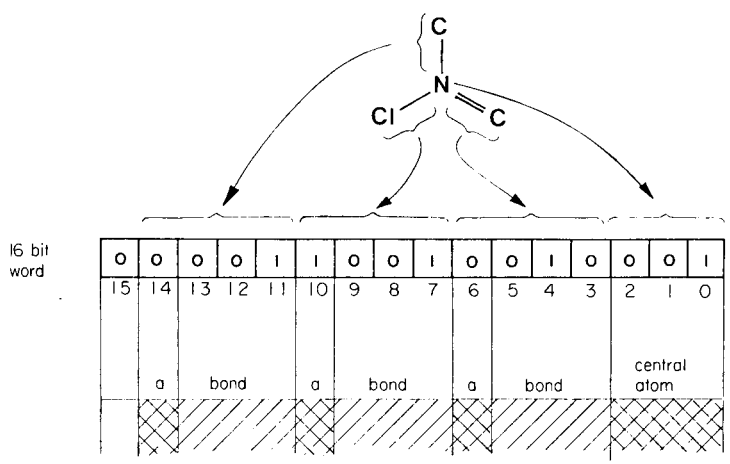


Fig. 4. Packing of a fragment into a 16-bit word. The sequence number of this fragment is 1 + 16 + 128 + 1024 + 2048 = 3217. The farthest left bit is set to 1 if there are more than three neighbours linked to the central atom. In the case that C=N—C atoms are taken from the ring, the bonds would be assigned numbers 5 and 4, respectively, and thus the sequence order of neighbours would change.

scheme can easily be expanded by using longer strings of bits, achieving greater diversity in fragment coding, and the numbering of fragments is canonical; moreover, the chemical meaning or image of the fragment can be retrieved directly from its number, without searching any tables and the address of a given fragment can easily be computed by using a hash algorithm. If the decomposition of the CT into fragments is done in order to create an inverted file of fragments for substructure searches, a prior procedure must also be used, i.e. the decomposition of larger fragments (three or more neighbours) into the entire set of smaller ones. This task must be carried out to ensure that any less complicated substructure created by the user will be retrieved from the stored structures. To clarify this statement, Fig. 5 shows the decomposition of hexafluorobenzene into two different fragments each of which appears six times in a complete decomposition. If the hexafluorobenzene structure is stored only in the inverted file under the two fragments shown, this structure will not be retrieved if the user tries to find all compounds containing the benzene ring, thus creating a simple ring as a query. The query structure (benzene ring in this case) will be decomposed into six equal fragments which will not match either of those from hexafluorobenzene. With this in mind, all larger fragments (containing three or more neighbours to the central atom) are decomposed to smaller subfragments.

When the 16-bit scheme is used for fragment encoding, theoretically more than 30 000 different fragments can be distinguished. Of these, slightly more than 1000 are chemically reasonable and, in fact, only about 300 different fragments were encountered in dealing with the data base of 1016 structures.

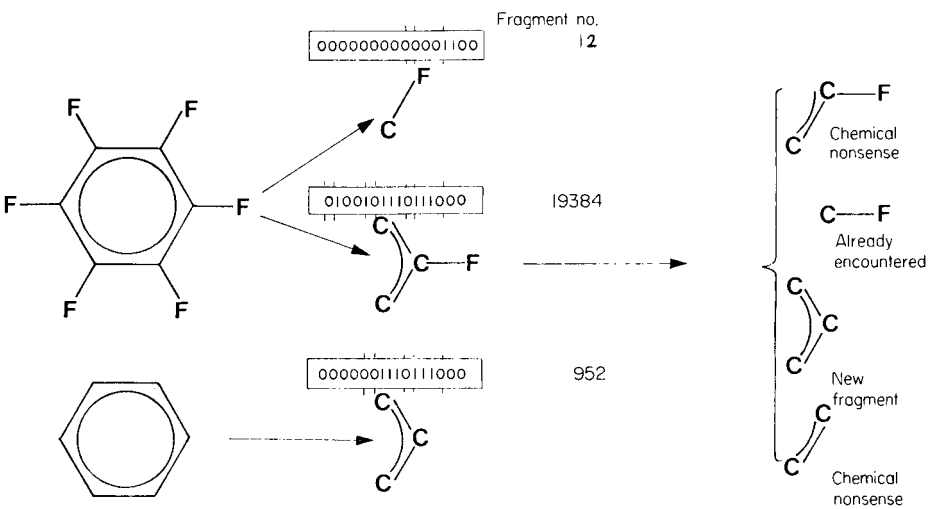


Fig. 5. Decomposition of hexafluorobenzene and benzene into fragments. Hexafluorobenzene is finally decomposed to three different fragments. The second step decomposition enables the user to retrieve this structure if the benzene ring is input as the query.

There is not much to say about the pure retrieval algorithms; the standard inverted file technique [5, 20] is used to retrieve the list of identity (ID) numbers of compounds having the desired spectral or structural features. Each new list is matched against the old one and the resulting common ID numbers are saved under the name specified by the user at the beginning of the search. If the old and new lists do not have common ID numbers, the old list is saved and the message is printed or displayed to the user. The user is able to see the remaining file, or issue any desired command at any step within the search. There can be as many as 30 files (lists of ID numbers) saved under different names and later recalled and inspected by the user. The search can be made either on one of the four compressed master files of 1016 items (structures, i.r., mass, or ^{13}C-n.m.r. spectra) or on any predefined list that was retrieved and stored during the session. To avoid very long records containing ID numbers of frequently occurring items, the starting entry address for each key (peak position or structural fragment) in the inverted files is hashed and, if necessary, chained by using the twin-prime number hash algorithm [5, 20].

Complete file manipulation, enabling logical operations among the files, has not yet been elaborated in detail and will be described at a later date.

The first impression of a system's ability and quality is normally given by the man—machine interface or by the flexibility and diversity of the command language and its interpretation. The merits of the system are often judged by the difficulties found by the user in applying the command language of the system or in appreciating the computer responses to the commands of the user (right or wrong). The response of the system to a wrong command (misspelling, wrong order of data, incorrect sequence of commands, etc.) is sometimes judged more severely than poor quality of retrieved results, possibly because some users seem to expect the system to "know" what they wanted to do (despite wrong commands), or because the user cannot comprehend the results displayed.

These aspects emphasize that much effort must be put into the design and construction of the command language. Even if it is implemented for CIS running on a minicomputer, diversity and flexibility must be built into the decoding algorithms for the interpretation of commands. In addition, a special file with condensed instructions on how to handle various troublesome events must be accessible during the session. In KISIK, this is achieved by using a special flag or pointer that changes its value according to the last correct command issued by the user. The flag is actually the starting address of the location on the SOS file where the appropriate message is stored.

The command language should normally be capable of accepting and interpreting strings of words, numbers, and separators such as, com_1, p_{111}, $p_{112} \cdots p_{11n}$, p_{121}, $p_{122} \cdots p_{12n}$, $\ldots com_2$, p_{211}, $p_{212} \cdots p_{21m}$, where com_i is any valid, predefined command directing the program to fulfil the specified step and

p_{ijk} is the kth parameter of the jth parameter set of the ith command. Each command may thus be accompanied by parameters explaining the step required in more detail. If parameters are absent but expected to be present, default values should be used. Most frequently, one command is associated with only one parameter set; however, for experienced users, repeated typing of the same command becomes tedious and faster communication is desirable. Table 1 shows two examples of command strings, one consisting of single step commands and the other using combined commands. Both command strings have the same effect.

The most important commands, together with explanations of parameters, are presented in Table 2. KISIK, as operated in this laboratory has all commands in Slovenian, but an English version for demonstration purposes has been installed in the same computer.

CONCLUSION

The chemical information system KISIK is an offspring of the batch-oriented combined retrieval system that was implemented on the CDC

TABLE 1

Example of two command strings leading to the same result. Underlined words are typed in by the user generating a query structure

Command string 1	Command string 2
Command: CHAIN,8 Command: IMAGE C1*C2*C3*C4*C5*C6*C7*C8 Command: BOND,2,3,2 Command: BOND,4,5,3 Command: CHAIN,1,7 Command: ATOM,8,F Command: ATOM,9,CL Command: IMAGE C1*C2+C3*C4#C5*C6*C7*F8 * C9 L	Command: CHAIN,8,IMAGE C1* C2* C3* C4* C5* C6* C7* C8 Command: BOND,2,3,2,4,5,3 Command: CHAIN,1,7 Command: ATOM,8,F,9,CL,IMAGE C1*C2+C3*C4#C5*C6*C7*F8 * C9 L

TABLE 2

List of the most important commands and parameters and their interpretation in two most important options of the CIS SEARCH and GENESIS.
(There are three other options, WLN, UPDATE and CALCULATOR.) These commands can be shortened, if the system can choose only one. For example, the shortest version of the command ATOM is A, while the command IDENTIFY must be written using at least two letters ID, to distinguish between IR and IMAGE. The commands and parameters are separated by commas.

Command	Parameters	Description
ATOM	*n*, *el*	Changing the type of atom with the number *n* into the *el*. Initially all atoms are assumed to be carbons
BOND	*n*1,*n*2,*b*	Changing the bond between the atoms *n*1 and *n*2 into *b*. Default values for all bonds are 1 and 4, for chains and rings, respectively
CHAIN	*n*,*n*1	Generation of the *n* atom long chain linked to the atom number *n*1 in the already existing query structure. If the command is given without the parameter *n*1 the chain is linked to the last-mentioned atom
CNMR		Start of the search through the ^{13}C-n.m.r. spectra file
CT		Display of the connection table and ring atoms description of the given structure
EXIT		Exit from any option but not from the CIS. All temporary files are saved and can be recalled when desired
FRAGMENTS		Decomposition of the structure to its fragments using which the substructure search will retrieve the required structures
IDENTIFY	*n*	Identification of the compound with the ID number *n*. The identification gives the chemical name and structural image, while molecular formula and i.r., m.s. or ^{13}C-n.m.r. spectra can be obtained on request
IMAGE		The two-dimensional structure image is displayed
IR	*MI*	Start of the search through the infrared file. If the parameter *MI* is added, the input positions are expected in μm, otherwise in cm^{-1}
MF		Molecular formula of the compound is displayed
MS		Start of the search through the file of mass spectra
NEW		Clearing the memory for new entries
RING	*n*	Generating a ring containing *n* carbon atoms. The command
RING	*n*,*n*1	can be issued without atom parameters *n*1 and *n*2 or with
RING	*n*,*n*1,*n*2	one only. In the first case the ring of *n* atoms is linked with a single bond to the last-mentioned atom. In the second case the ring is formed in such a way that atom *n*1 is included within the ring, and if both parameters *n*1 and *n*2 are typed in, the ring will contain both. This last way is particularly useful for the construction of fused rings
SEARCH		Start of the substructure search using the generated structure as a query
SHOW	*xxxx*	List of ID's stored on the file named *xxxx* is displayed together with the chemical names of compounds
STATISTIC	*xxxx*	The frequency of fragment appearance in all structures from the file *xxxx* is calculated and displayed
SOS		Condensed instructions of commands to overcome troubles. It can be invoked at any step
WLN		Input of the query structure using Wiswesser Line-Formula Chemical Notation

Cyber 172 computer [21]. Despite many demonstrations to other analytical chemists and industrial staff, the earlier system evoked little enthusiasm, mainly because of the remote mainframe computer and bad connection facilities from laboratories, but also because of the reluctance to use batch-oriented computers. Demonstrations of the interactive and minicomputer-oriented KISIK version were better received. Many more potential users have a minicomputer in their laboratories than have the possibility of connection to a mainframe computer. Another attractive feature is the possibility of creating collections which exactly fit specific needs.

The present version of KISIK requires 28K words of central memory on PDP 11/34 which calls for at least 48K of memory. As indicated above, the programs and data banks for 1 016 compounds require 1.8 Mbyte of space and can be stored on a 2.5-M byte disk (RK05J). However, few of the potential users have three different spectrometers, and the system data bank can be easily expanded to over 3000 compounds described by only one type of spectrum within the given disk space.

It seems probable that the continuing increase in the number of users of large and efficient CIS's (like the NIH—EPA CIS) will be accompanied by increasing development of smaller dedicated information systems. The words "small" and "dedicated" refer to the data banks rather than to the program packages. The use of files which are of modest size but rich in information will give results more quickly and more reliably than is possible from larger systems. This is by no means a challenge to the large file-oriented CIS's that will continue to provide the major support for general and inter-disciplinary basic research in chemical and borderline areas. The role of the minicomputer CIS's should be complementary to the larger mainframe systems.

The financial support of the Research Community of SR Slovenia is gratefully acknowledged.

REFERENCES

1 Extended Abstracts, I, International Conference on Computer-based Analytical Chemistry, Portorož, September 1979, Vestn. Slov. Kem. Drus., 26, Supplement (1979).
2 S. R. Heller, G. W. A. Milne and R. J. Feldman, Science, 195 (1977) 253.
3 J. E. Dubois, in Computer Representation and Manipulation of Chemical Information, W. T. Wipke, S. R. Heller, R. J. Feldmann and E. Hyde (Eds.), John Wiley, New York, 1974, p. 239.
4 L. A. Gribov, M. E. Elyashberg and M. M. Raikhshtat, J. Mol. Struct., 53 (1979) 81.
5 J. Zupan, Anal. Chim. Acta, 103 (1978) 273.
6 J. T. Clerc and J. Zupan, Pure Appl. Chem., 49 (1977) 1827.
7 J. Zupan, M. Penca, M. Razinger, B. Barlič and D. Hadži, Anal. Lett., 12(A2) (1979) 109.
8 Infrared Spectroscopy, Its Use in the Coating Industry (740 IR Spectra), Fed. Soc. Paint Technology, 1969.
9 D. O. Hummel and F. Scholl, Atlas der Kunstoff-Analyse, Hochpolymere und Harze (1800 IR Spectra), Carl Hauser, Munich, 1968.

10 D. O. Hummel and F. Scholl, Atlas der Kunststoff-Analyse, Zusatzstoffe und Verarbeitungshilfmittel (1100 IR Spectra), Carl Hauser, Munich, 1973.
11 D. F. Johnson and W. C. Jankowski, ^{13}Carbon NMR Specra (500 ^{13}C n.m.r. Spectra), John Wiley, New York, 1972.
12 Sadtler Collection, Philadelphia, Monomers and Polymers (700 IR Spectra).
13 M. Razinger, J. Zupan, M. Penca and B. Barlič, Interactive Simulation of Infrared, Mass and ^{13}C-n.m.r. Spectra, J. Chem. Inf. Comp. Sci. (1980) in press.
14 H. L. Morgan, J. Chem. Soc., 5 (1965) 107.
15 E. G. Smith and P. A. Baker, The Wiswesser Line-Formula Chemical Notation, 3rd edn., CIMI, Cherry Hill, NJ, 1975.
16 J. E. Dubois, in The Chemical Applications of Graph Theory, A. T. Balban (Ed.), Academic Press, New York, 1976.
17 K. K. Agrawal, Transformation and Canonization Algorithms for Graph Representable Structures with Applications to Heuristic Program for the Synthesis of Organic Molecules, Thesis, State University of New York at Stony Brook, Tech. Rep. 63, 1976.
18 Z. Hippe, R. Hippe, G. Kruczek and M. Dec, Coding of Chemical Structures in the CONOL-II Notation, Report, I, Lukasiewicz Technical University, Rzesow, Poland, 1979.
19 J. Zupan, The WLN-to-CT Conversion Program, to be published.
20 D. E. Knuth, The Art of Computer Programming, Vol. III, Addison—Wesley, Reading, 2nd edn., 1975, pp. 389, 506.
21 J. Zupan, M. Penca, D. Hadži and M. Marsel, Anal. Chem., 49 (1977) 2141.

Analytica Chimica Acta, 122 (1980) 117—138
Computer Techniques and Optimization

THE NIH-EPA CHEMICAL INFORMATION SYSTEM IN SUPPORT OF STRUCTURE ELUCIDATION†

STEPHEN R. HELLER*
Environmental Protection Agency, Washington, DC 20460 (U.S.A.)
GEORGE W. A. MILNE
National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20205 (U.S.A.)
(Received 25th September 1979)

SUMMARY

Progress in the development of the NIH—EPA chemical information system is reviewed.

A major activity in modern chemistry is the identification of chemical substances from spectral measurements. Whatever measurement technique is used, the task generally devolves into one of recognising a 'fingerprint' given by the unknown in, e.g., a mass spectrometer, when thousands of 'fingerprints', e.g., mass spectra, of known compounds are available. The NIH—EPA Chemical Information System permits 'fingerprint recognition' in a variety of efficient and inexpensive ways and is used very heavily by scientists all over the world.

A much more challenging task now dominating CIS development is the prediction of the properties of a substance from its molecular structure. The short-term promise of such predictive ability is a tremendous saving in resources; large numbers of expensive and time-consuming measurements can be obviated by strategies in which the properties of all substances in a set can be predicted, based on a few selected experimental measurements. As a longer-term goal of the CIS development, an understanding of the relationships between structure and properties is beginning to flow from studies facilitated by access to the large, evaluated, high-quality numeric data bases contained in the CIS.

This presentation will center on the first of the above capabilities of the CIS, namely the use of the CIS spectral data bases for structure elucidation.

The quantity of chemical data has been expanding in recent decades, but until the advent of third-generation (integrated circuitry) computers, handling and using this vast amount of information has presented insuperable problems. With modern computer technology, the cost of computation has come down, while accessibility has increased through the use of computer net-

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

works, accessible over standard telephone lines. In these circumstances, a highly interactive, disk-oriented chemical information system of numerical data has been developed which is readily and inexpensively available to Agency laboratories, contractors, grantees, scientific collaborators and the general public.

The early computer developments in search systems minimized the high cost of mass storage by maintaining data files on magnetic tape rather than drums or disks. Data bases can be stored very inexpensively on tape, but can only be searched sequentially because tape is not susceptible to random access. Magnetic disks are random access devices with considerable storage capacity. Because of this, data stored on disks can be searched very rapidly. Until recently, the costs of disks, controllers and the other items necessary to use such equipment has precluded their use for large data bases. These costs have now decreased markedly and since the use of disks for data storage permits interactive computing, the Chemical Information System uses disk exclusively for the storage of data.

Interactive computing is a significantly different process from batch, off-line, or batch on-line computing and a different philosophy must be used in the design of programs for such work if the system is to take advantage of the power of disk storage and access. A major problem in searching a chemical data base is that the best questions are often unknown. An interactive system can provide the answer to a question immediately and this will enable the user to see the deficiencies in the question and to frame a new query. In this way, a feedback loop can be built, in which the scientist acts as a transducer, 'tuning' the query until the system reports precisely what is required. The NIH—EPA Chemical Information System, described here in some detail, has been designed around this general approach. The building of a general system requires that all of the known information be gathered together within one computer system and that all this information be of acceptable accuracy and precision. This is an ideal that the CIS has tried to reach by using the best scientific data available from the areas of science which are felt most likely to be of value in solving the diverse problems of improving health and the environment.

SYSTEM DESIGN

The NIH—EPA CIS consists of a collection of chemical data bases, together with a library of computer programs for interactive searching through these disk-stored data bases. In addition, the CIS has a data-base referral capability as well as a data analysis software system. It can be thought of as having four main areas: numerical data bases; data analysis software; structure and nomenclature search system, and data base referral.

The numerical data bases that are part of the CIS include files of mass spectra [2], ^{13}carbon nuclear magnetic resonance [3], x-ray diffraction data for crystals [4] and powders [5, 6], mammalian acute toxicity data [7],

and aquatic toxicity data [8]. There are bibliographic data bases associated with the mass spectrometry, x-ray crystallography and n.m.r. areas and these have been included within the CIS [9]. The analytical programs include a family of statistical analyses and mathematical modelling algorithms [10] and programs for the second-order analysis of n.m.r. spectra [11] and energy minimization of conformational structures [12]. Programs that design chemical synthesis are being tested and may, if viable, become part of the CIS in the future [13].

The center or hub of the CIS is the Structure and Nomenclature Search System (SANSS) [14], which allows the user to search through data bases of structures (such as those associated with collections of mass spectra) for occurrences of a specific structure or substructure. With this program, for example, regulatory agencies considering the problem of collecting data on aromatic *m*-dichloro compounds could proceed as follows: The substructure search shown in Fig. 1 could be conducted and this would find all occur-

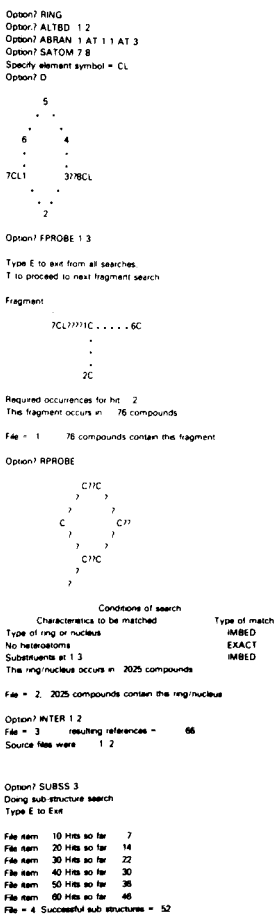


Fig. 1. Search for aromatic chloro, bromo compounds in the CIS Unified Data Base.

rences of *m*-dichloro compounds in the 41 data bases searched. In turn, by reference to the Toxic Substances Control Act (TSCA) and International Trade Commission (ITC) lists, or the Resource Conservation and Recovery Act (RCRA) list, etc., this could lead to information such as the number of chemicals involved, the dollar volume of chemicals affected, and so on. If necessary, a subset of these chemicals could be defined and investigated in further detail.

In the area of structure elucidation, if one had evidence that an unknown contained a particular substructure, a search might reveal that there were n.m.r. spectra to compare with such a similar structure, but no i.r. spectra, suggesting that a n.m.r. spectrum would be more useful than an i.r. spectrum in attempts to identify the unknown. As more and more data bases are collected and merged into the SANSS, it becomes a catalog of files of chemicals [15]. Recently the structure of the SANSS files was reorganized so that this referral capability has become much more efficient, by using an integrated data base of the 41 files shown in Fig. 2.

The entire CIS structure can be viewed (Fig. 3) as a network of independent numerical data bases, linked together through the SANSS hub, by using the Chemical Abstracts Service Registry Number (CAS REGN) as the unique universal chemical identifier for each compound. The use of the CAS REGN to tag all CIS files, was codified in EPA regulation 2800.2 in 1975 [16]. With the passing of the TSCA in 1976, the use of the CAS REGN was extended to the TSCA inventory and this establishes the link between regulatory data and scientific data, both within the CIS and in the literature. In Fig. 3, the solid circles represent systems running on commercial systems. The solid boxes represent systems which are currently being put through their final testing at NIH (to smooth any rough edges) before they are considered operational and placed in the commercial system. The dotted triangles are systems under development. Lastly, the dotted lines to the circles refer to operational systems to which the CIS can link, but that are on other computers on the Telenet network (Telenet Communications Corp., Vienna, VA 22180), which is the telecommunications network used by the CIS.

1 – TSCA Candidate List (33,568)
2 – CIS Mass Spectrometry (31,834)
3 – CIS Carbon 13 NMR Spectrometry (7,901)
4 – EPA Pesticides – Active Ingredients (1,452)
5 – EPA OHM/TADS (858)
6 – Cambridge Xray Crystal (14,677)
7 – Merck Index (8,981)
8 – EPA Pesticides – Analytical Ref. Stnds. (473)
9 – EPA STORET (234)
10 – EPA Chemical Spills (577)
11 – EPA AEROS SOTDAT (572)
12 – NIMH Psychotropic Drugs (2,036)
13 – EPA AEROS SAROAD (65)
14 – NBS Proton Affinities (439)
15 – CPSC CHEMRIC (890)
16 – EPA Pesticides – Registered Inert Ingredients (734)
17 – NBS Gaseous Ions (3,163)
18 – NFDA Hazardous Chemicals (396)
19 – FDA/EPA Pesticides Ref. Standards (613)
21 – U.S. International Trade Commission (9,188)
22 – NBS Xray Crystal (18,229)
25 – EPA Effluent Guidelines (119)
26 – EPA Organic Chemical Producers (375)
27 – IPC Chemical Product (104)
28 – IPC Chemical Plant (103)
29 – NSF Chemicals List (225)
30 – EROICA Thermodynamics (4,488)
31 – PHS-149 Carcinogenic Activity (4,437)
32 – NIOSH RTECS (20,336)
33 – NIOSH NOHS (4,559)
35 – ORNL EMIC (4,034)
36 – ORNL ETIC (3,244)
43 – EPA – Selected Organic Air Pollutants (573)
45 – Clean Air Act – Section 112 (5)
58 – EPA/NCTR Study (1976) (85)
59 – EPA Environmental Carcinogen Assessment Program (26)
66 – EPA – Restricted Use Pesticides (22)
67 – EPA – Compounds For Mutagenicity Evaluation (25)
70 – CIIT Priority Chemicals Lists (Toxicological) (26)
77 – NLM CHEMLINE (26,392)
82 – NMFS Survey Of Trace Elements (15)

Fig. 2. List of the current 41 collections which comprise the CIS Unified Data Base.

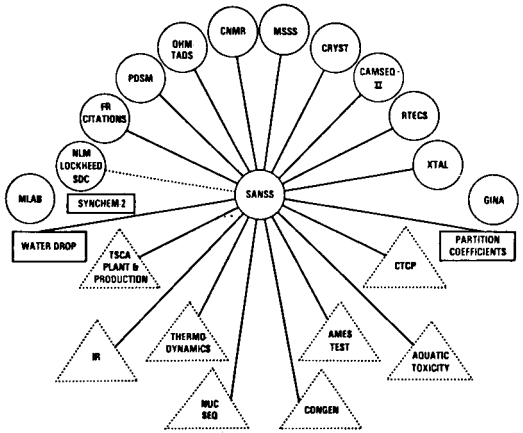


Fig. 3. The CIS components, their status and how they are linked together: (—) directly linked on same computer; (---) linked via a network and CAS REGN. For further explanation, see text. CAMSEQ-II — Conformational Analysis of Molecules in Solution by Empirical & Quantum-Mechanical Techniques; CNMR — Carbon-13 Nuclear Magnetic Resonance Spectral Search System; CONGEN — Constrained Generation of Structures from Molecular Formulas; CRYST — Cambridge Crystal Data Base Search System; CTCP — Clinical Toxicity of Commercial Products; FR — Federal Register; GINA — Graphical Interactive NMR Analysis; IR — Infra Red Spectral Search System; MLAB — Mathematical Modelling Laboratory; MSSS — Mass Spectral Search System; NLM — National Library of Medicine; NUCSEQ — Nucleotide Sequence Data Base; OHM-TADS — Oil & Hazardous Materials Technical Assistance Data System; PDSM — Powder Diffraction Search Match System; RTECS — Registry of Toxic Effects of Chemical Substances; SANSS — Structure and Nomenclature Search System; SDC — System Development Corporation; SYNCHEM-2 — Synthesis Design Program; TSCA — Toxic Substances Control Act; WATER DROP — Distribution Register of Organic Water Pollutants; XTAL — Single Crystal Search System.

CIS SYSTEM DEVELOPMENT

A general protocol for updating of CIS components or the addition to the CIS of new components has been established. A schematic diagram is shown in Fig. 4. In the first phase, a data base is acquired from one of a variety of sources. Some of the CIS data bases have been developed specifically for the CIS, an example of this being the mass spectral data base [2]. Others, such as the Cambridge Crystal File [4], are leased for use in the CIS and still others, such as the x-ray powder diffraction file [5], are operated within the CIS by their owners, in this case the Joint Committee on Powder Diffraction Standards. In other cases, the information comes from other Government Agencies which retain responsibility for the file, its contents and its maintenance. An example of such a file is the NIOSH RTECS.

If the data base is to be made searchable, some reformatting, sorting and inversion of files is usually required and this is carried out on the NIH IBM 370-168, which is well-suited to processing large files of data. Once inverted lists have been prepared, they are transferred to the NIH PDP-10 computer

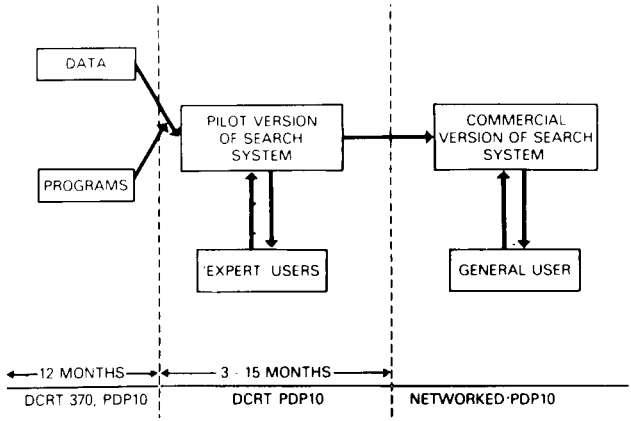


Fig. 4. Protocol for adding a component to the CIS.

which is primarily a time-sharing computer, and the programs for generating the searchable files and for searching through these files are written. Analytical, data base-independent programs of the CIS are usually written entirely on the PDP-10.

From this work, there finally emerges a pilot version of each CIS component. Access to this pilot version on the NIH PDP-10 is provided to a small number of people who can log into the NIH computer by telephone, using long distance calls if necessary. These users are provided with free use of the test component and in return, they test it thoroughly for errors and deficiencies. Problems are reported to the development team, which attempts to deal with them. Depending upon the size and complexity of the component, this testing phase can last as long as 18 months.

When testing is completed, the entire component is exported to a networked PDP-10 in the private sector and the version on the NIH computer is no longer maintained. The component in the private sector is available to the general scientific community, including Government Agencies, and is used on a fee-for-service basis. In this phase, the system is then maintained by a Government contractor, but efforts are made to generate sufficient interest in, and use of the system so that other organizations will take over the responsibility of maintenance.

COMPUTER FACILITIES USED BY THE CIS

Programs of the CIS have usually been designed for use with a DEC PDP-10 computer system, because the PDP-10 is one of the better time-sharing systems available and has been adopted by a number of commercial computer network companies as the main vehicle for their networks. Transfer of a program from the NIH PDP-10 to a network PDP-10 is usually fairly straightforward, and use of a networked computer is favored because the alternative

philosophy of exporting programs and data bases to locally operated PDP-10 computers is less workable and contains a number of deficiencies that are overcome by a network. Most important among these is the fact that use of a networked machine means that data bases need only be stored once, at the center of the network. A great deal of money is thus saved because duplicate storage is not necessary. Further, a single copy of a data base is easy to maintain, whereas updating a data base that resides on many computers is virtually impossible. Finally, communications between systems personnel and users is very simple in a network environment, as is monitoring of system performance.

For these and other reasons, the policy of disseminating the CIS via a networked PDP-10 computer was adopted at the outset and has proved to be quite successful. A typical American network of this sort has something under 100 nodes, i.e. local telephone access is available in about 100 locations. These are mainly in the U.S., but a substantial number may be found in other countries. Further, some computer networks are now interfaced to the Telex network, thus making their computer systems available worldwide. Costs are reasonable. The only equipment that is required to establish access to a computer network, is a telephone-coupled computer terminal that will operate at about 300 baud (30 characters/second). A cathode ray terminal capable of running at 1200 baud can be used.

COMPONENTS OF THE CIS

Mass spectral search system (MSSS)

The Mass Spectral Search System (MSSS) is the oldest component of the CIS. The first version of MSSS was developed in 1971, and the system has been used as a prototype for more recently designed components. Developed as a joint effort between NIH, EPA, NBS (National Bureau of Standards) and the Mass Spectrometry Data Centre (MSDC) in England, the current MSSS data base contains about 33 900 mass spectra representing the same number of compounds. This has been derived from an archival file containing some 60 000 spectra of the same 33 900 compounds [17]. Computer techniques have been employed to assign every spectrum a quality index [18] and where duplicate spectra appear in the archive file, the best spectrum is selected for use in the working file. All compounds in the archive have been assigned a Chemical Abstracts Service (CAS) registry number, a unique identifier that is used to locate duplicate entries for the compounds, find the compound in other CIS files and provide structure and capabilities for looking up synonyms throughout the CIS.

Searches through the MSSS data base can be carried out in a number of ways. With the mass spectrum of an unknown in hand, the search can be conducted interactively (Fig. 5). In this search the user finds that 89 data base spectra have a peak (minimum intensity 60%, maximum intensity 100%) at an m/z value of 272. When this subset is examined for spectra containing

```
NIH/EPA/MSDC Mass Spectral Search System (Version 3.8), July 1979 Edition (1.2)

Latest news for MSSS
 15 Aug 79: MSSS Data Base Updated --32,191 Spectra Now Available
OPTION: PEAK

TYPE PEAK,MIN INT,MAX INT
CR TO EXIT, 1 FOR ID#,REGN,QI,MW,MF AND NAME

USER:272,60,100
 # REFS   M/E PEAKS
    89     272

NEXT REQUEST: 237,10,70
 # REFS   M/E PEAKS
     6     272,237

NEXT REQUEST: 357,5,30
 # REFS   M/E PEAKS
     1     272,237,357

NEXT REQUEST: 1
  ID#      REGN  QI  MW  MF
29819     143500 728 486 C10CL10O

                                    NAME
                                    1,3,4-METH
END-2H-CYCLOBUTA[CD]PENTALEN-2
-ONE, 1,1A,3,3A,4,5,5A,5B,6-
DECACHLOROOCTAHYDRO- (8CI9CI)

CHLORDECONE
CLORDECONE
COMPOUND 1189
DECACHLOROKETONE
DECACHLOROPENTACYCLO[5.2.1.02,
6.03,9.05,8]DECAN-4-ONE
ENT-16391
GC 1189
KEPONE
MEREX
```

Fig. 5. PEAK search in the MSSS.

a peak at m/z 237 with intensity of between 10 and 70%, only 6 spectra are found. The entering of a third peak, at m/z value of 357 (with an intensity between 5 and 30%), narrows the search to one answer, which is then printed out. In the example shown, the answer, Kepone, is shown with a number of synonyms used in naming this chemical, as well as other identifying information. If there had still been a large number of answers after entering the three peaks used in this example, the search could have been reduced further to a manageable number of spectra by entering further peaks. In addition, the data base can be examined for all occurrences of a specific molecular weight or a partial or complete molecular formula. Combinations of these properties can also be used in searches. Thus all compounds containing, for example, five chlorines and whose mass spectra have a base peak at a particular m/z value can be identified.

In contrast to these interactive searches, which are of little appeal to those with large numbers of searches to carry out, there are available two batch-type searches which accept the complete spectrum of the unknown and examine all spectra in the file sequentially to find the best fits. These are the KB (forward search) and PBM (reverse search) algorithms. Spectra can be entered from a teletype, but in a more powerful approach, a user's data system can be connected to the network for this purpose and the unknown spectra down-loaded into the network computer for searching. Once an identification has been made, and the name and registry number of the data base compound are reported to the user, the data base spectrum can be listed or, if a CRT terminal is being used, plotted, to facilitate direct comparison of the unknown and standard spectra.

Also within the MSSS are the accumulated files of the Mass Spectrometry Bulletin (UKCIS, Nottingham, England) which contains about 60 000 citations to papers published since 1967 on mass spectrometry, and may be searched interactively for all papers by given authors, and all papers dealing

with one or more specific subjects or with one or more particular elements. Citations dealing with general index terms may also be retrieved. The interactive nature of the search provides great control to the user, as indicated earlier [19].

No numerical codes are used by the system. A search for a specific subject can be carried out by entering the subject word itself. If the word 'mass' is entered, searches for 7 terms (all those containing the fragment 'mas' i.e. mass spectra, mass discrimination, mass measurement, etc.) are conducted and the user is asked to select the one of interest. In this way, precise knowledge of the correct subject words or of their correct spelling is not necessary.

With the current high level of interest in chemical ionization mass spectrometry, there is a need for a reliable file of gas-phase proton affinities. The task of gathering and evaluating all published gas-phase proton affinities was completed by Hartmann et al. [20]. This file, which has about 400 critically evaluated affinities drawn from the open literature, can be searched on the basis of compound type or the proton affinity value.

The MSSS has been widely available through computer networks since 1971 and is currently resident upon the Interactive Sciences Corporation (ISC) computer.

^{13}Carbon nuclear magnetic resonance (CNMR) spectral search system

The data base that is used in the CNMR search system consists currently of 8 700 ^{13}C-n.m.r. spectra. As in the case of the MSSS, every compound has a CAS registry number, and exact duplicate spectra have been removed from the file. The CNMR file is still small but is growing at a fairly steady rate and should benefit considerably from recent international agreements that all major compilations of such data will, in future, be pooled [21]. Searching through this data base, as in the case of the MSSS, can be interactive or not [3]. In the interactive search, the user enters a shift, with an acceptable deviation. The algorithm reports the number of file spectra fitting this criterion. The names of the compounds whose spectra have been retrieved can be listed, or the list can be reduced by the entry of a second chemical shift. A search for spectra of compounds having a specific molecular formula is also possible, but there is no capability for searching on molecular weight, a parameter of little relevance in this field.

If an interactive search is not appropriate to the problem, a batch-type search through the data base is available by using the techniques described by Clerc et al. [22]. To institute such a search, the user enters all the chemical shifts from the unknown. The entire unknown spectrum is compared to every entry in the file and the best fits are reported. This program searches for the absence of peaks in a given region as well as for the presence of peaks and is thus capable of finding compounds structurally similar to the unknown material.

When a search is completed, the user is provided with the CAS Registry numbers of compounds with spectra that fit the input data. The names of

these compounds are also given. If more information is required, the complete entry for a given CAS registry number can be retrieved. This includes a numbered structural formula, the name, molecular formula and registry number of the compound, experimental data pertaining to the spectrum and the entire spectrum, together with single frequency off-resonance decoupled multiplicities and, for 60% of the spectra, relative line intensities and assignments.

In addition to the library aspects of CNMR, an interface between CNMR and SANSS (the structure search component of the CIS) has been written. This allows a user to define a substructure and then examine the chemical shifts associated with particular carbon atoms of interest. The shift data is plotted out to the user, with appropriate standard deviations for the data, which should be helpful in structure elucidation problems.

X-ray crystallographic search system (CRYST)

This is a series of search programs working against the Cambridge Crystal File [4], a data base of some 22 000 compounds for which full atomic coordinate data are available, and over 25 000 bibliographic entries dealing with published crystallographic data, mainly for organic compounds. The entry for each compound contains the compound name, its molecular weight and CAS Registry number, the space group in which it crystallizes and the parameters of the unit cell of the crystals as well as the atomic coordinate data. The file may be searched on the basis of any of these parameters as shown in Fig. 6, which outlines a search for any compounds that crystallize in space group *P* 1 and have molecular weights between 250 and 300. As can be seen, there are 133 entries with the correct space group (temporary file 1) and 2 038 with molecular weight between 250 and 300 (temporary file 2). The intersection of these files reveals that only 21 compounds (temporary file 3) meet both specifications, and the first of these compounds, crystal sequence number 849, is listed. All the compounds in this file have been registered by the CAS and these data are currently being merged into the CRYST system along with the connection tables for the structures. This data base is therefore searchable on a structural or substructural basis, as are all the other files of the CIS. Once an entry of interest in the Cambridge x-ray file has been located by one of the search programs, its 'crystal sequence number' can be used to retrieve the appropriate literature reference, structure, or atomic coordinate data.

This data base possesses complete literature references to all entries in the file [4]. This information has been made the basis of a system for searching the literature pertaining to the x-ray diffraction study of organic molecules, by author(s), title words, journal, year, etc., in a manner similar to the Mass Spectrometry Bulletin Search System. The system generates temporary files from searches, as in the SANSS, and files can be intersected on request with 'AND' or 'NOT' operators. Once a paper of interest has been identified, all

```
XRAY CRYSTALLOGRAPHY SEARCH SYSTEM (09-MAR-78)

CRYST TEMPORARY FILES WILL BE ANN. TMP AND ARNN. TMP

30-June-1978 CRYST status information

FILE NOW CAMBRIDGE CRYSTAL SUMMER 77
OPTION: SPGR

SPACE GROUP SYMBOL
>P1

FILE = 1 REFERENCES = 133 ITEM = P 1
OPTION: SMOLS 1

TYPE MOLECULE WEIGHT RANGE

>250.300

FILE = 2 MERGED REFERENCES = 2038

OPTION: INTER 1 2

FILE = 3 INTERSECTED REFERENCES = 21
SOURCE FILES WERE: 1 2
OPTION: SSHOW 3

START WITH NTH REFERENCE (1) = 1

SHOW EVERY NTH REFERENCE (1) = 1

STRUCTURE 1 CRYSTAL SEQUENCE 849
```

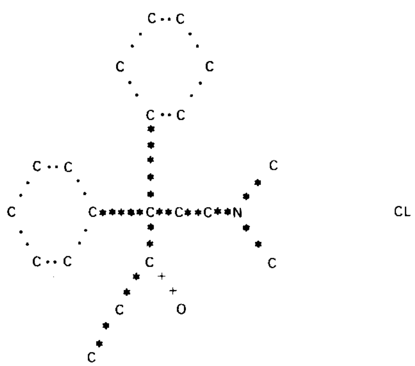


Fig. 6. Space group and molecular weight search in the Cambridge crystal data base.

the crystallographic information in that paper can be examined because the crystal sequence serial number associated with the paper can be used in the crystallographic search system. Alternatively, the CAS Registry number of any particular compound can be used to retrieve any data of interest on that compound from other files of the CIS.

X-ray crystal data search system (XTAL)

The National Bureau of Standards (NBS) has collected a file of data pertaining to some 45 000 crystalline materials, including those in the Cambridge file described above. The data in the NBS file (Tape 9, National Technical Information Service, Springfield, VA 22151) include the cell parameters, the number of molecules in the unit cell, the measured and calculated densities of the crystal and two determinative ratios, such as A/B and A/C (but no coordinate data). Every compound in the file is identified by its name, molecular formula and CAS Registry number and the file can be structurally

searched by the CIS structure and nomenclature search system as described below.

Searches through this data base for crystals with specific space groups or densities have been developed and it is possible to locate crystals with reduced cells of given dimensions. It is hoped that this will prove to be a very rapid method of identifying compounds from the readily measured crystal properties.

X-ray powder diffraction search match (PDSM) system

A collection of powder diffraction patterns is a very effective means by which to identify materials; indeed, one of the earliest search systems in chemical analysis was based upon such data by Hanawalt et al. [23] forty years ago. The importance of these data in TSCA can be seen by examining the TSCA Inventory regulations for treatment of confidential chemicals [24]. Section 710.7 of these regulations indicates that EPA intends to rely on powder diffraction data to assure the validity and seriousness of a manufacturer's request for treating information on a chemical as confidential.

The data base of some 33 000 powder diffraction patterns that is used in the CIS [5] is a direct descendent of that with which Hanawalt et al. carried out their pioneering work [23]. A problem that arises in connection with this particular component stems from the fact that powders are frequently mixtures of different crystalline phases, and so the patterns obtained experimentally are often combinations of one or more file entries. A reverse searching program [6], that examines the experimental data to see if each entry from the file is contained in it, has been written after the general approach of Abramson [25], and seems to cope with this particular difficulty. A subtraction routine to help in identifying mixtures has also been implemented.

NIOSH RTECS search system (RTECS)

The National Institute for Occupational Safety and Health (NIOSH), is required by law to prepare a list of all the toxic effects of chemicals that can be found [26]. The Registry of Toxic Effects of Chemical Substances (RTECS) is the data base created and updated annually by NIOSH to comply with this law. In 1979 the data base consisted of some 35 000 chemicals and the toxicity associated with each of these chemicals. The NIOSH RTECS is the first non-spectroscopic CIS data base and has proven to be a very valuable addition. Interest in the data base has been shown by many groups within EPA involved in the implementation of TSCA. For example, work is now underway to link spectral data with the NIOSH toxicity data so that as a result of a mass spectral identification, the appropriate authorities can quickly be informed if the chemical identified is toxic and hence requires immediate action. The RTECS data base can be searched in a number of ways, including NIOSH number, CAS Registry number, type of animal tested, route of dosage, LD50, LCLO, etc. The file is also linked to the SANSS so

that structure-activity relationships may be examined.

An example of a NIOSH RTECS search is shown in Fig. 7, which presents a search for all rodent oral LD50 toxicity data with values less than 1 mg kg^{-1}; the system indicates there are 50, and one is listed.

Structure and nomenclature search system (SANSS)

The CAS registry number is a unique identifier for a compound, and may be in a CIS data base, then, armed with its CAS Registry number, the user all the synonyms that the CAS has identified for the compound, in addition to the name used in the CAS 9th Collective Index. Further, the registry number can be used to locate in the CAS files, the connection table for the compound structure. This is a two-dimensional record of all atoms in the molecule together with the atoms to which each is bonded and the nature of the bonds [27] and is the basis of the substructure search component of the CIS.

The purpose of the SANSS is to permit a search for a user-defined structure or substructure through data bases of the CIS. If a substructure is found to be in a CIS data base, then, armed with its CAS Registry number, the user can access that file and locate the compound and hence retrieve whatever data are available for it. There are several ways of searching the CIS Unified Data Base. The main ones are: Name/Fragment Name Search (NPROBE); Nucleus/Ring Search (RPROBE); Fragment Search (FPROBE); Structure Code Search (SPROBE), Molecular Weight, Molecular Formula, Partial Formula; Total Atom-by-Atom, Bond-by-Bond Search (SUBSS) and Total or Full Structure Search (IDENT).

While structure searching is very important and cannot be replaced by other methods (such as fragment searching, linear notations or name searching), the ability to search for a chemical by name or partial name (NPROBE), is very useful in many cases. In particular, if one wishes to search for a drug or pesticide, many of which have simple and short trivial names, a name search is likely to be the best method because such compounds are often complex cyclic structures, difficult to draw. In the example shown in Fig. 8, a name search is conducted for the carcinogen TCDD. The program is asked (using the SSHOW command) to print out the files in which this one chemical containing the name fragment TCDD appears, along with its molecular formula, structural diagram and correct Chemical Abstracts Index name, as well as the synonyms associated with the chemical.

As the first step in a sub-structure search, the user must define the sub-structure of interest to the computer. This is done with a family of structure generation programs which can, for example, create a ring of a given size, a chain of a given length, a fused ring system and so on. Branches, bonds and atoms can be added and the nature of bonds and atoms can be specified. In the absence of definition, an atom is presumed to be carbon. As the query structure is developed using these commands, the computer stores the growing connection table. If the user wishes to view the current structure

You are now in the RTECS system (Version 2.7-April, 1979)

Latest news for RTECS . . .
6 July 79; TSHOW Option Prompting Message Changed
Option? SEARCH

Animal type? ROD

Dosage method? ORL

Type of measure? LD50

Do you want all values of toxicity? NO

Upper numeric limit and units? 1 MG/KG

Lower numeric limit and units? 0 MG/KG

Search of the RTECS data base for:
Animal = ROD
dosage method = ORL
measure = LD50
and toxic effect = ALL
between the limits 1 MG/KG and 0 MG/KG
yields 50 entries on file 3

Option? TSHOW 3

How many (E to Exit)? 1

CAS number = 62-74-8 NIOSH number = AH9100000

ORL-HMN	LDLO:	714	UG/KG	TFX:	34ZIAG	-,542,69
ORL-HMN	LDLO:	5	MG/KG	TFX:	27ZTAP	3,71,69
UNK-MAN	LDLO:	5	MG/KG	TFX:	AJPEAG	36,1427,46
ORL-RAT	LD50:	220	UG/KG	TFX:	PHRPA6	61,672,46
IPR-RAT	LD50:	800	UG/KG	TFX:	JAPMA8	36,59,47
SCU-RAT	LD50:	5	MG/KG	TFX:	JAPMA8	36,59,47
ORL-MUS	LD50:	4	MG/KG	TFX:	JAPMA8	36,59,47
IPR-MUS	LD50:	15	MG/KG	TFX:	JOCEAH	23,1567,58
ORL-DOG	LD50:	66	UG/KG	TFX:	JPETAB	101,82,50
IVN-MKY	LD50:	5	MG/KG	TFX:	AJPEAG	36,1427,46
IPR-CAT	LD50:	300	UG/KG	TFX:	AJPEAG	36,1427,46
SCU-RBT	LD50:	281	UG/KG	TFX:	JPETAB	95,62,49
IVN-RBT	LD95:	500	UG/KG	TFX:	PAREAQ	1,383,49
ORL-GPG	LD66:	400	UG/KG	TFX:	JAPMA8	36,59,47
IHL-GPG	LC50:	100	MG/M3	TFX:	JCSOA9	-,1773,48
IPR-GPG	LD50:	378	UG/KG	TFX:	JPETAB	95,62,49
ORL-PGN	LD50:	4	MG/KG	TFX:	TXAPA9	20,57,71
ORL-CKN	LD50:	5	MG/KG	TFX:	JAPMA8	36,59,47
ORL-QAL	LD50:	18	MG/KG	TFX:	TXAPA9	20,57,71
ORL-SQL	LD50:	300	UG/KG	TFX:	JAPMA8	36,59,47
IPR-SQL	LD50:	400	UG/KG	TFX:	JAPMA8	36,59,47
ORL-DCK	LD50:	4810	UG/KG	TFX:	TXAPA9	22,556,72
SCU-FRG	LD50:	1000	MG/KG	TFX:	AJPEAG	36,1427,46
IPR-DOM	LD50:	200	UG/KG	TFX:	AJPEAG	36,1427,46
ORL-DOM	LD50:	250	UG/KG	TFX:	AJURAH	9,370,48
IMS-DOM	LD50:	700	UG/KG	TFX:	AJPEAG	36,1427,46
UNK-MAM	LD50:	2	MG/KG	TFX:	AMIHAB	14,178,56
ORL-BRD	LD50:	5	MG/KG	TFX:	JAPMA8	36,59,47

There are review articles available
There are standards and regulations that apply
for this chemical
ACETIC ACID, FLUORO-, SODIUM SALT
C2-H2-F-O2 .NA

Fig. 7. Search for acute toxicity data.

```
Option? NPROBE

Fragment or whole name search (F/W) (F)?W

Specify name (CR to exit) TCDD

File 5, 1 compounds having name TCDD
Specify name (CR to exit)

Option? SSHOW 5

Structure 1 CAS Registry number 1746-01-6
CIS Mass Spectrometry
Cambridge Xray Crystal 1746 01 6 01
EPA Effluent Guidelines
NIOSH RTECS HP35000
ORNL ETIC
NLM CHEMLINE TOXLINE, MEDLINE
      C    O    C                    C12H4C14O2
     . .  * *  . .
    .  . *  * .  .
CLXC    C    C    CXXCL
    .   .    .    .
CLXC    C    C    CXXCL
    .  . *  * .  .
     . .  * *  . .
      C    O    C

Dibenzo[b,e][1,4]dioxin, 2,3,7,8-tetrachloro (9CI)
Dibenzo-p-dioxin, 2,3,7,8-tetrachloro- (8CI)
Dioxin (herbicide contaminant)
TCDBD
TCDD
2,3,7,8-Tetrachlorodibenzo-p dioxin
2,3,7,8-Tetrachlorodibenzo-1,4-dioxin
```

Fig. 8. Name Search (NPROBE) for TCDD.

at any point, the display command (D) can be invoked. This command, using the current connection table, generates a structure diagram similar to those in Figs. 9—11. This can be printed at a conventional terminal and Figs. 9—11 were so printed.

When the appropriate query structure has been defined, a number of search options can be used to find occurrences of this query structure in the data base. The two most useful search options are the fragment probe and the ring probe. The fragment probe will search through the assembled connection tables of the data base for all occurrences of a particular atom-centered

```
OPTION? RING
OPTION? ALTBD 1 2
OPTION? ABRAN 1 AT 1 1 AT 4
OPTION? SATOM 7
SPECIFY ELEMENT SYMBOL = SE
OPTION? FPROB 1

TYPE E TO EXIT FROM ALL SEARCHES,
T TO PROCEED TO NEXT FRAGMENT SEARCH

FRAGMENT:
          7SE?????C . . . . 6C
                    .
                    .
                    .
                   2C

REQUIRED OCCURRENCES FOR HIT : 1
THIS FRAGMENT OCCURS IN 57 COMPOUNDS

FILE = 1, 57 COMPOUNDS CONTAIN THIS FRAGMENT
```

Fig. 9. SANSS Fragment probe search.

```
Option? D

        7CL
        ?
       ?
5?????4
?      ?
?      ?
10     3??6CL
 ?    ?
  ? ?
   2

Option? EXIM 3 4
Optron? RPROBE

O?????C
 ?     ?
 ?     ?
C      C??
 ?     ?
  ? ?
   C
   ?
   ?
```

Conditions of search

Characteristics to be matched	Type of match
Type of ring or nucleus	EXACT
Heteroatoms at 1	EXACT
Heteroatoms are O	IMBED
Substituents at 4 3	IMBED

This ring/nucleus occurs in 979 compounds

File = 7, 979 compounds contain this ring/nucleus

Option? SSHOW 7

How many (E to Exit)? 1

Structure 1 CAS Registry number 50-81-7
TSCA Candidate List: R000-6313
CIS Mass Spectrometry
CIS Carbon 13 NMR Spectrometry: 50-81-7.01 TO 50-81-7.02
Cambridge Xray Crystal: 50-81-7.01 TO 50-81-7.02
Merck Index: 0855
EPA Pesticides—Registered Inert Ingredients
U.S. International Trade Commission
NBS Xray Crystal: 50-81-7.01 TO 50-81-7.02
NSF Chemicals List: 211
PHS-149 Carcinogenic Activity: A1169
NIOSH RTECS: CI76500
NIOSH NOHS: M0462,80147
ORNL EMIC
ORNL ETIC
NLM CHEMLINE: TOXLINE

C6H806

```
                O
                +
               +
          O*****C
          *     *
          *     *
O**C**C**C      C**O
       *    *   +
       *    * +
       O     C
             *
             *
             *O
```

L-Ascorbic acid (8CI9CI)
Adenex
Allercorb
Antiscorbic vitamin
Antiscorbutic vitamin
Ascorbajen
Ascorbic acid
Ascorbutina
Ascorin
Ascorteal
68 more names available

Fig. 10. SANSS Ring/Nucleus probe search.

```
Option? D

   10CL  70
   ?     +
   ?     +
8CL3??1?? 2P?50?11
   ?  ?  ?
   ?  ?  ?
   9CL40 60
         ?
         ?
         12

Option? IDENT

Total proton count for this structure is
    (P for program estimate) P

Total proton count based upon normal conditions is 8
Are there any abnormal valence or charge conditions which
    would affect this count (Y/N)? N

Proton count for node 2 (D to display structure)? O

File 8, This structure is contained in 1 compounds.

Option? SSHOW 8

Structure 1 CAS Registry number 52-68-6
TSCA Candidate List: R001-5032
CIS Mass Spectrometry
EPA Pesticides—Active Ingredients: 57901
EPA OHM/TADS: 72T16519
Cambridge Xray Crystal: 52-68-6.01
Merck Index: 9303
EPA Pesticides—Analytical Ref. Stnds.: 6780
EPA Chemical Spills
FDA/EPA Pesticides Ref. Standards: 48
PHS-149 Carcinogenic Activity: C0147
NIOSH RTECS: TA07000
ORNL EMIC
ORNL ETIC
NLM CHEMLINE: TOXLINE, MEDLINE

C4H8C1304P

    CL      O
    *       +
    *       +
CL*C**C**P**O**C
    *   *   *
    *   *   *
    CL  O   O
            *
            *
            C
```

Phosphonic acid, (2,2,2-trichloro-1-hydroxyethyl)-, dimethyl ester (8C I9CI)
Agroforotox
Anthon
Bayer L 13/59
Chlorofos
Chloroftalm
Chlorophos
Chlorophthalm
Chloroxyphos
Combot
44 more names available

Fig. 11. Complete structure (IDENT) search.

fragment, i.e. a specific atom, together with all its neighbors and bonds. The user may specify particular fragments which are thought to be fairly unique and characteristic of the query structure. Alternatively, a search for every fragment in the query structure may be requested. A fragment probe is shown in Fig. 9. The query structure contains only one relatively unique node, C1, and this is the one which is sought in the data base. It is found to occur 57 times and a temporary file of just those particular entries is stored as file 1. This can be accessed either for the purpose of listing its contents, or, as is shown, for entry into SUBSS.

The ring probe search is a search for all structures in the data base containing the same ring or ring system as the query structure. A ring that is considered to be an answer to such a query must be the same size as that in the query structure. It must also contain at least as many heteroatoms as the query structure, but the nature of the heteroatoms can be required by the user to be the same as or different from, that in the query structure. The type of bonding is not considered in an RPROBE search. Thus with a query structure of furan, the only 'exact' answer is furan but the user may permit the retrieval of other answers including, tetrahydrofuran and thiophene. An example of a ring probe search is given in Fig. 10. Here the query structure is a 3,4-dichlorofuran, but imbedded matches for heteroatom type and substituents have been allowed and so the list of 979 answers will include any disubstituted pyrroles as well as any disubstituted furans and so on. A higher degree of substitution may also be permitted.

In addition to these structural searches, there are a number of 'special properties' searches that are often useful as a means of reducing a large list of answers resulting from structure searches. The special-properties searches include searches for a specific molecular weight or range of molecular weights and a search for compounds containing a given number of rings of a given size. Searches may also be conducted for the molecular formula corresponding to the query structure, or for other user-defined molecular formulae. This may be specified completely or partially and the number of atoms of any element may be entered exactly or as a permissible range.

If the purpose is to determine only the presence or absence in a data base of a specific structure, this can be accomplished with the search option IDENT (Fig. 11). This program hash-encodes the query structure connection table and searches through a file of hash-encoded connection tables for an exact match. The search, which is very fast (with respect to both CPU and elapsed time) by substructure search standards, has been designed specifically for those users who, to comply with the Toxic Substances Control Act [28], have to determine the presence or absence of specific compounds in Environmental Protection Agency files.

If one has completed ring-probe and fragment-probe searches for a specific query structure and is still confronted with a sizeable file of compounds that satisfy the criteria that were nominated, a substructure search through this file may be carried out. This involves an atom-by-atom, bond-by-bond com-

parison of every structure and will retrieve any compound which contains the query structure imbedded in it, as can be seen from Fig. 1.

Finally, a link between the CIS structure search system (SANSS) and the vast scientific literature has been established using the Lockheed DIALOG system. The link consists of using an intelligent terminal (with local memory) to store CAS Registry numbers found in a CIS SANSS search for later transmission to the Lockheed system after the user logs off CIS and (since Lockheed and CIS are both on Telenet) logs into DIALOG without having to re-dial the phone. This semi-automatic interfacing of different computers holds considerable promise for future linking of computers throughout the world.

The structure and nomenclature search system is the center of the CIS and operates on a unified data base of 41 files (Fig. 2). The SANSS data bases are in the process of being updated with an additional 55 files, and a further 100 files are now being processed. The 55-file update, which will bring the number of files in SANSS to 95, is scheduled for the spring of 1980. The whole system is available for general use on the ISC computer.

N.m.r. graphical interactive spectrum analysis (GINA)

Many proton n.m.r. spectra can be satisfactorily analyzed by hand, and such first-order analysis is then quite satisfactory for assigning chemical shifts and coupling constants to the various nuclei involved. In certain cases, however, second-order effects become important and as a result, more or fewer spectral lines than are indicated by first-order considerations will result. A way to analyse such spectra is to estimate the various coupling constants and chemical shifts and then, using any of a variety of standard computer programs [11], calculate the theoretical spectrum corresponding to these values. The calculated spectrum can be compared to the observed spectrum and a new estimate of the data can be made. In this way, by a series of successive approximations, the correct coupling constants and chemical shifts can be determined.

The CIS component GINA is based on the programs developed by Johannesen et al. [29], and permits these operations in real time in an interactive fashion. The program is designed for use with a vector cathode-ray tube terminal upon which each new theoretical spectrum can be displayed for comparison by the user with the observed spectrum.

Mathematical Modelling System (MLAB)

MLAB is a program set developed by Knott and Shrager [10] which can assimilate a file of experimental data, such as a titration curve, for example, and perform on it any of a wide variety of mathematical operations. Included amongst these are differential and integral calculus, statistical analysis (mean and standard deviation, curve and distribution fitting and linear and non-linear regression analysis). Output data can be presented in any form, but the PDP-10 program is especially powerful in the area of graphical output.

Data can be displayed as two- or three-dimensional plots which can be viewed and modified on a CRT terminal prior to photography.

Conformational analysis of molecules in solution (CAMSEQ-II)

A problem of long standing in chemistry has been to estimate the relationship between the conformation of a molecule in the crystal, as measured by x-ray methods, with that in solution where barriers to rotation are greatly reduced. A sophisticated program set for Conformational Analysis of Molecules in Solution by Empirical and Quantum-mechanical methods (CAMSEQ-II) has been developed for this purpose by Hopfinger and co-workers [12]. This program can run in batch or interactively. As input data, it requires the structure of the compound and this can be provided as a set of coordinate data from x-ray measurements. Alternatively, it can be entered interactively in the form of a connection table or the program can simply be provided with a CAS registry number, and if the corresponding connection table is in the files of the CIS, it will use that.

The first task is to generate the coordinate data corresponding to a particular compound. Then the free energy of this conformation in solution is calculated. Next the program begins to change torsion angles specified by the user in the conformation and with each new conformation, a statistical thermodynamic probability is calculated, based upon potential (steric, electrostatic, and torsional) functions and terms for the free energy associated with hydrogen-bonding, molecule-solvent and molecule-dipole interactions.

WaterDROP

Over the past few years, the improved sensitivity of analytical methods, particularly mass spectrometry, has permitted the accumulation of information about environmental pollutants. In the water systems of the United States, the EPA has found many chemicals, of which over 1300 have been identified. Similar results have been obtained in Europe, under the guidance of the European Economic Community (EEC). The result of these activities has been the accumulation of considerable information about the identities of potentially toxic chemicals and where they may have been found. An EPA research laboratory (EPA, ERL, Athens, GA 30601) realizing the need for a centralized source for collection, storing and disseminating this information has started to develop a Distribution Register of Organic Pollutants in Water (WaterDROP). The WaterDROP system contains the identity of the chemical found, the sampling site and date, reporting laboratory, analytical method used and date of entry into the system. The data will be collected in a number of ways, but the main automatic collection of data for the system is expected to come from laboratories within EPA, by interaction of local g.c.-m.s.-computer—internal report systems, with the results fed to a central computer and thence to a centralized reporting file. In this system, MSSS users identify the unknown toxic pollutant by a KB-type search. The Biemann search procedure will be modified so that EPA laboratories are

required to enter additional information when conducting a search. As each laboratory identifies an unknown, the central computer builds up a data base of information for WaterDROP. The results of all these searches will be a centralized report file, such as shown in Table 1. With international cooperation in building this data base, the WaterDROP file should grow quickly.

The data bank will be published by the EPA, as well as being searchable under the SANSS and the WaterDROP software that has been developed. The system is being tested at NIH; when it is networked, it will be possible to answer questions about the localities where a given chemical (or class of chemicals) is found, patterns of distribution of chemicals in water indicating problems with plant effluents, etc. Answers to these and other similar questions, coupled with toxicity data from RTECS and other CIS sources, should provide valuable technical information for regulation and control of pollutants.

Aquatic toxicity (AQUATOX)

Owing to the importance of fish in human nutrition, concern over the danger posed to fish by chemicals is being recognized as a major activity of EPA and other U.S. Government groups. A data bank of aquatic toxicity is being developed by EPA, in conjunction with ASTM Committee E-35.21.01. This data bank, expected to be available for testing on CIS shortly, will have information on the chemicals found in fish, reported toxicities, literature citations, common and scientific names of the species studied, temperature, pH and hardness or salinity of the water in the study, salinity of the water and a comments section for other desired information related to the study.

There are other data bases of valuable numerical information which are being built and obtained or expanded from existing sources. These include

TABLE 1

Sample entries for the WaterDROP system from a modified MSSS Biemann search

Registry	Compound name	River	River mile	Long.	Lat.	Date	S.I. Value	Labo tory
62759	Methanamine, *N*-methyl- *N*-nitroso-dimethylamine, *N*-nitroso-dimethylnitroso-amine DMN DMNA	Ohio	137	85.32	40.05	04-23-77	0.981	173
80626	2-Propenoic acid, 2-methyl, methyl ester methacrylic acid, methyl ester methyl methacrylate MME methyl 2-methylpropenoate	Hudson	097	72.20	40.80	04-24-77	0.978	110

files of i.r. spectra, mutagenesis and teratogenesis studies, partition coefficients and thermodynamic data. Further information on all aspects of these projects is available from the authors.

Conclusion

One of the first goals of the CIS was to produce a series of searchable chemical data bases for use by working chemists with no special computer expertise. A second aim was to link these data bases together so that the user need not be restricted to consideration of a single type of data. The various problems inherent in these plans included acquisition of data bases, design of programs, dissemination of the resulting system and linking, via CAS registration numbers, of the various CIS components. These problems have been solved conceptually and, to a large extent, practically. It is now possible therefore to review the system in an effort to define future goals.

Searches through more than one data base in combination would be very desirable. For example, one often possesses both mass spectral and n.m.r. data for an unknown and it would be useful to be able to identify any compounds that match these data in a single search. Work is going on in this area to interface programs so that this approach can be tested.

In another development under discussion, it is hoped that the CONGEN program developed for the DENDRAL project [30] will soon be merged into CIS within the next year. This program, which generates structures corresponding to a specific empirical formula, could be extremely useful in a strategy for structure-solving based on the CIS. It is easy to envisage situations in which a reduced set of structures could be produced for consideration by CONGEN. Confirmation for any of them could then be sought in the spectral data bases, given the registry number.

In a different approach, the power of pattern recognition techniques could be assessed within some of the very large files contained in the CIS. This could be useful because there is little reported work of this sort on large files. The value of such methods in handling the problem of identification of true unknowns such as water pollutants is under study. Programs designed to test mass spectra for the presence in the compound of elements or groups, such as halogens and aromatic rings, have been written [31] and their utility as pre-filters on mass spectral data prior to data base searching will be tested.

REFERENCES

1 S. R. Heller, G. W. A. Milne and R. J. Feldmann, Science, 195 (1977) 253.
2 S. R. Heller, H. M. Fales and G. W. A. Milne, Org. Mass Spectrom., 7 (1973) 107; S. R. Heller, D. A. Koniver, H. M. Fales and G. W. A. Milne, Anal. Chem., 46 (1974) 947; S. R. Heller, R. J. Feldmann, H. M. Fales and G. W. A. Milne, J. Chem. Doc., 13 (1973) 130; S. R. Heller and G. W. A. Milne, J. Chem. Info. Comp. Sci., 16 (1976) 176.
3 D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, Org. Magn. Reson., 11 (1978) 535.

4 O. Kennard, D. G. Watson and W. G. Town, J. Chem. Doc., 12 (1972) 14.
5 G. McCarthy and G. G. Johnson, paper C3, Proceedings of the American Crystallographic Association meeting, State College, PA, 1974.
6 R. G. Marquart, I. Katsnelson, G. W. A. Milne, S. R. Heller, G. G. Johnson Jr. and R. Jenkins, Applied Cryst., 12 (1979) 629.
7 NIOSH, Registry of Toxic Effects of Chemical Substances, Vols. 1 and 2, DHEW (NIOSH) 78-104-A, GPO 017-033-0027101, Government Printing Office, Washington, DC, 1977.
8 C. Stephan, EPA, Duluth, MN 55804, unpublished data.
9 These include the bibliographic file associated with the data in ref. 4, and the Mass Spectrometry Literature Bulletin, published by the Mass Spectrometry Data Centre, UKCIS, The University, Nottingham, England.
10 G. D. Knott and R. I. Shrager, Assn. Comp. Machin., SIGGRAPH Notes 6 (1972) 138.
11 S. R. Heller and A. E. Jacobson, Anal. Chem., 44 (1972) 2219.
12 H. J. R. Weintraub and A. J. Hopfinger, Int. J. Quant. Chem., 9 (1975) 203; R. Potenzone, E. Cavicchi, H. J. R. Weintraub and A. J. Hopfinger, Comp. Chem., 1 (1977) 187.
13 H. L. Gelernter, A. F. Sanders, D. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spitzer and J. E. Searleman, Science, 197 (1977) 1041.
14 R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller and B. Koch, J. Chem. Info. Comp. Sci., 17 (1977) 157; G. W. A. Milne, S. R. Heller, A. E. Fein, E. F. Frees, R. G. Marquart, J. A. McGill, J. A. Miller and D. S. Spiers, J. Chem. Info. Comp. Sci., 18 (1978) 185.
15 S. R. Heller, G. W. A. Milne and R. J. Feldmann, J. Chem. Info. Comp. Sci., 16 (1976) 232.
16 EPA Order 2800.2, May 27, 1975.
17 NIH—EPA—MSDC data base, National Bureau of Standards, Office of Standard Reference Data, Washington, DC 20234.
18 D. D. Speck, R. Venhataraghavan, F. W. McLafferty, Org. Magn. Reson., 13 (1978) 208.
19 V. A. Vinton, G. W. A. Milne and S. R. Heller, Anal. Chim. Acta, 95 (1977) 41.
20 K. Hartmann, S. Lias, P. J. Ausloss and H. M. Rosenstock, Publication NBSIR 76-1061, July 1976.
21 C. L. Citroen, Netherlands Information Combine, 2600 AA, Delft, The Netherlands.
22 J. T. Clerc, R. Schwarzenbach, J. Meili and H. Koenitzer, Org. Magn. Reson., 8 (1976) 11.
23 J. D. Hanawalt, H. W. Rinn and L. K. Frevel, Ind. Eng. Chem., 10 (1938) 457.
24 Environmental Protection Agency, Toxic Substances Control Act (TSCA) Inventory reporting Requirements, Federal Register, 42,247, December 23, 1977, 64572-64596.
25 F. P. Abramson, Anal. Chem., 47 (1975) 45.
26 PL 91-596, Occupational Safety and Health Act of 1970 (OSHA), section 20(a).
27 L. J. O'Korn, in R. E. Christoffersen (Ed.), Algorithms for Chemical Computations, ACS Symposium Series 46, 1977.
28 PL-94-469, Toxic Substances Control Act, 1976.
29 R. B. Johannesen, J. A. Ferretti and R. K. Harris, J. Magn. Reson., 3 (1970) 84.
30 R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, J. Am. Chem. Soc., 97 (1975) 5755.
31 W. Meisel, M. Jolley, S. R. Heller and G. W. A. Milne, Anal. Chim. Acta., 112 (1979) 407.

Analytica Chimica Acta, 122 (1980) 139—149
Computer Techniques and Optimization

EXTRACTION OF INFORMATION ON THE CHEMICAL STRUCTURE OF MONOFUNCTIONAL COMPOUNDS FROM RETENTION DATA IN GAS—LIQUID CHROMATOGRAPHY BY PATTERN RECOGNITION METHODS†

J. F. K. HUBER* and G. REICH

Institute of Analytical Chemistry, University of Vienna, Waehringer Straße 38, A-1090 Vienna (Austria)

(Received 21st December 1978)

SUMMARY

A method for the determination of structural features of unknown compounds, based on multi-dimensional gas—liquid chromatographic retention data, is presented. The species investigated are monofunctional compounds. The retention index is used as retention parameter. A two-step classification procedure is reported. The first step is the determination of a correction parameter for the retention index, the skeleton number. With these correction terms, the retention data are modified, and the second classification step for the determination of functional groups is executed with the modified data. The best classification obtained is based on the linear learning machine method. A 10-dimensional data set, i.e. the use of 10 stationary liquids, is sufficient for the total classification.

Gas—liquid chromatography (g.c.) is used primarily in the qualitative and quantitative analysis of volatile compounds. Substances are identified by use of a library of retention data, mostly in the form of indices. The method proceeds by a simple numerical comparison of the measured retention values and the data stored in the library. Problems with this method arise if an unknown compound is not contained in the library of reference compounds; under these circumstances no correct answer can be found. Furthermore, the certainty of identification is limited by the precision and accuracy of measurement.

It is obvious that information about the chemical structure of a given compound is contained in a set of retention values, measured on different stationary liquid phases. The identification of at least a few structural features from these data must be possible by applying appropriate methods. The encoding of the structural information is quite complex, however, and no simple algorithm can be expected to solve this problem. In this paper, an attempt is made to solve the classification of compounds according to

†This paper was presented at the International Conference on Computers and Optimization in Analytical Chemistry, Amsterdam, April 1978.

their functional groups from multi-dimensional gas chromatographic retention data by the use of pattern recognition methods.

Several papers have been published in which pattern recognition methods are used in analytical chemistry. Fundamental aspects [1—7], and applications in mass spectroscopy [8—17] or infrared spectroscopy [18—22] have been discussed but few papers have dealt with applications in chromatography. The subjects of these papers are the systematics of stationary liquid phases [23—26], and the identification of complex mixtures (petroleum samples) by g.c. using the peak positions in a chromatogram as dimensions and the peak heights as coordinates of a point, and classifying this point by means of pattern recognition methods [27].

EXPERIMENTAL

The calculations were performed by means of a large computer (Control Data CYBER 73) with a core memory of 69K words of 60-bit, disk and tape storage. A process computer (Digital Equipment PDP-15) with a core memory of 8K words of 18-bit, dual DEC-tape, line printer and a modem, was used for interactive job development and data input. The two computers were connected through a dial telephone line. The software used on the PDP-15 comprised a few utility programs for job editing and remote job entry. The software on the CYBER 73 consisted of programs for development and management of the data base. The main operating tool was ARTHUR, a program system for complex multidimensional data analysis by pattern recognition methods [28].

The values for the data base were taken from the literature [29], where the relative retentions and the retention indices of 367 compounds on 77 stationary liquid phases at two temperatures are listed. In the present work the retention indices were used. The largest data matrix is formed by the values on 74 stationary liquid phases at a temperature of 120°C. By exclusion of all compounds with missing values, the data of 142 compounds remain. After a feature reduction procedure has been applied, a data matrix of 19[illegible] compounds on 10 liquid phases is obtained. The feature selection procedure used will be described later. The composition of the data set with respect to functional groups is given in Table 1. The part of the Wiswesser Line Notation (WLN) describing the functional group is used to indicate the structural characteristics of a given category of compounds. These symbols will be used in the following tables to indicate the categories.

RESULTS AND DISCUSSION

The classification for functional groups with the original data set, as expected, was not very successful, according to a fundamental principle of pattern recognition, because the points (i.e. compounds) with similar properties, i.e. same functional groups, must form a group close together in the

TABLE 1

Type and number of compounds to be classified

Class of compounds	WLN	Number of compounds	Class of compounds	WLN	Number of compounds
Alcohols	Q	55	Esters	VO	60
Aldehydes	VH	16	Aromatic hydrocarbons	R	9
Ketones	V	22	Alkanes and haloalkanes	—	12
Ethers	O	25			

n-dimensional space. This is not true for this data set as can be seen in Fig. 1, which shows the eigenvector projection of homologous compounds, represented by the first two eigenvectors. If homologous compounds were similar in their retention behavior they should form small clusters. Actually, they form straight or curved lines in agreement with the finding that the retention index of the members of a homologous series increases linearly with the number of carbon atoms.

In order to achieve a closer grouping of the data of a homologous series, a pre-processing method has to be introduced. This method must be simple, requiring no additional information about unknown compounds. Such a pre-processing method was developed. It is based on the determination of the "skeleton number".

Development of the pre-processing procedure

In gas chromatographic structure elucidation, similar patterns of multidimensional retention data should be obtained for homologous compounds. The retention indices of homologous compounds, however, are not very similar because each CH_2—group gives an increment of 100 retention index

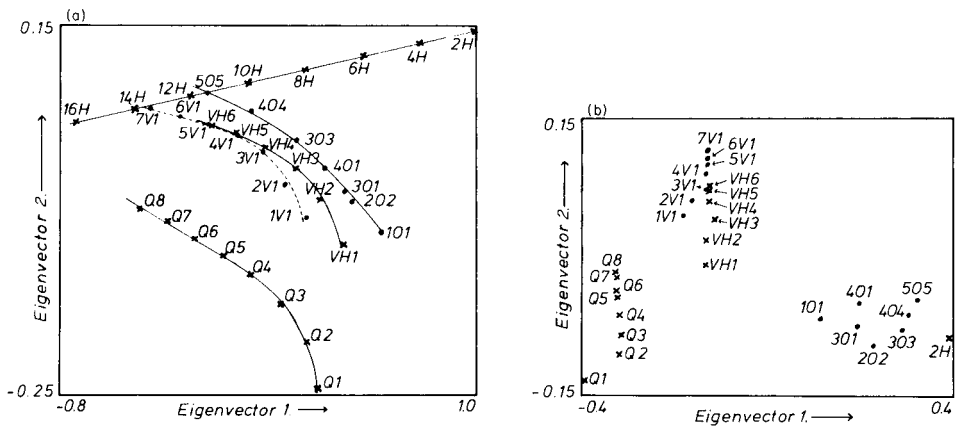


Fig. 1. Eigenvector projection of the retention indices of homologous series of n-alkanols (Qn), methyl-n-alkylketones (nV1), n-aldehydes (VHn), n-ethers (nOm) and n-alkanes (nH): (a) original data (I_{Ri}); (b) pre-processed data ($\Delta_s I_{Ri}$).

units. Therefore, suitably adjusted data have to be derived from the retention indices. It is necessary to find suitable correction terms for the retention indices and to calculate adjusted data before classification for functional groups can be started. The pre-processing was done with the data set of 142 compounds on 74 stationary liquid phases at 120°C. The linear learning machine method was applied for the classification according to functional groups.

The most obvious pre-processing was based on subtraction of the retention index, 100 n, of the preceding n-alkane, n, from the retention index, I_{Ri}, of a compound, i: $\Delta_n I_{Ri} = I_{Ri} - 100\,n$. However, the classification for functional groups with the modified data, $\Delta_n I_{Ri}$, was even worse than that from the original data, I_{Ri}. An alternative pre-processing involved subtraction of the retention index, 100 n_C, of the n-alkane with the same carbon number, n_C, from the retention index, I_{Ri}, of the compound, i: $\Delta_C I_{Ri} = I_{Ri} - 100 n_C$. This form of pre-processing gave significantly improved but not completely satisfactory results in the classification according to functional groups.

Another form of pre-processing followed from the observation that a branched alkane has a smaller retention index than the n-alkane with the same carbon number. Accordingly, another correction term was defined by the retention index of an n-alkane for which the number of carbon atoms is equal to the number, n_u, of unbranched (primary and secondary) carbon atoms of the compound, i, to be classified. Thus: $\Delta_u I_{Ri} = I_{Ri} - 100\,n_u$. This pre-processing gave slightly worse results than the previous approach. The next method was an extension of the preceding one. The correction term is defined by the sum of the number, n_u, of the unbranched carbon atoms and the number, n_h, of the hetero atoms without hydrogen: $\Delta_{u+h} I_{Ri} = I_{Ri} - 100\,(n_u + n_h)$. The results with this form of pre-processing were slightly better, but still not entirely satisfactory.

The final form of pre-processing incorporated all atoms except hydrogen. The correction term is given by the retention index of an n-alkane with a number of carbon atoms equal to the number of all atoms, n_s, of the compound, i, to be classified, excluding hydrogen. The modified data are obtained according to $\Delta_s I_{Ri} = I_{Ri} - 100\,n_s$. For a compound with the empirical formula $C_5H_{10}O_2$, for example, the skeleton number, n_s, is 7 and the correction term 700. This pre-processing procedure gave the best results in the classification for functional groups by pattern recognition, and was used in the further work.

Determination of unknown skeleton number

The skeleton number is established by using the linear learning machine [1, 2, 31, 32]. The patterns are separated into two categories by a linear hyperplane computed in a training procedure. The data of 199 compounds on 10 stationary liquids form the training set. Each compound in the training set is assigned a value, the skeleton number, which is defined by the number of atoms in each compound, excluding the hydrogen atoms. The compounds

are separated into two categories according to the following principle. The hyperplanes must be trained to separate compounds with close skeleton numbers. Therefore the following two categories are formed: all compounds with skeleton numbers 1, 2 and 3 constitute category 1, and all compounds with skeleton number 4, 5, 6, 7, 8, 9, 10, 11, 12 or 13 constitute category 2. The corresponding hyperplane is then determined. In the next training process, the compounds with skeleton number 4 are transferred from category 2 to category 1 and the hyperplane is again calculated; this separates the compounds with skeleton numbers 1—4 from compounds with skeleton numbers 5—13. These procedures are repeated, until 7 hyperplanes which separate 7 pairs of categories have been trained. The final two categories consist of the compounds with skeleton numbers 1—9 and 10—13, respectively.

This training process is quite lengthy. It can be shortened by excluding in each step one skeleton number from the training process, defining it as the threshold value in the separation process. The categories for the first training process resulting from this principle are: all compounds with skeleton numbers 1—3 in category 1 and all compounds with skeleton numbers 5—13 in category 2. The compounds with skeleton number 4 are excluded. For the second training process, the compounds with skeleton number 4 are added to category 1, and all compounds with skeleton number 5 are removed from category 2. Finally, 7 hyperplanes are again trained; the last hyperplane separates categories consisting of compounds with skeleton numbers 1—9 from those with numbers 11—13.

The skeleton number of an unknown compound can easily be calculated by multiplying its pattern vector by the decision vectors (hyperplanes). The resulting series of products change their signs at one place in the series. The corresponding threshold skeleton number is the calculated skeleton number of the unknown compound. This procedure is a modification of a multicategory learning machine [33, 34]. The results of these classifications are summarized in Table 2. Because of the limited data set, and the restriction from the classification algorithm [35—38] an adequate test set was not available to prove the classification. Therefore the investigations were limited to establishing the possibility of separation. The results are presented as the number of iteration cycles necessary to reach 100% separation. Because the number of iteration cycles also depends on the initialization of the decision vector, each classification was done twice. The two initialization hyperplanes were selected at right angles to each other. The results given are the average of both classifications. The heading $(n_s + 1)$ indicates the method with all compounds included; the heading $(n_s + 2)$ indicates the results of the method in which compounds with the threshold skeleton number $n_s + 1$ are excluded.

Feature reduction

In order to reduce the computing time, the amount of data must be minimized, therefore the minimum number of stationary liquids required for the

TABLE 2

Classification for the skeleton number, n_s, with the linear learning machine method

Lower category (LC) $1-n_s$	Upper category (UC) $(n_s + 1)-13$		Upper category (UC) $(n_s + 2)-13$	
	Number of compounds LC/UC	Number of cycles	Number of compounds LC/UC	Number of cycles
2	5/194	516	5/183	110
3	16/183	764	16/166	433
4	33/166	2978	33/134	547
5	65/134	1503	65/93	66
6	106/93	18896	106/61	757
7	138/61	1340	138/35	194
8	164/35	721	164/15	535
9	184/15	1051	184/5	73

classification was determined. The modified complete data set for 142 compounds and 74 stationary liquid phases at 120°C was categorized according to functional groups and the stationary liquids were weighted and selected by means of two weighting functions (Fisher and variance weighting). Classification for functional groups with the preprocessed data set was found to be possible with a minimum of 4 liquid phases (Table 3). The original data, I_R, of 142 compounds on the 15 most selected stationary liquids were then categorized for skeleton number and the stationary liquids were weighted and selected again. The classification for skeleton number was found to need a minimum of 10 liquid phases (Table 3). The rational choice of the stationary phases in multi-dimensional gas chromatography by pattern recognition methods will be discussed more fully in a future paper [39].

Evaluation of different classification methods

The basis for the classification according to functional groups is the data set which results from the pre-processing of the retention indices by subtracting the corresponding skeleton numbers times 100. If an eigenvector projection of this modified data set for homologous compounds is considered (Fig. 1b), it can be seen clearly that small clusters are far better achieved with the modified data than with the original data. The modified data show a significantly better clustering for given functional groups.

The same ten liquid phases selected for the determination of the skeleton number were then used for the classification according to functional groups. Four classification methods were compared: the K-nearest neighbor, the hierarchal clustering, the minimum spanning tree, and the linear learning machine methods. All four methods were applied to both the original data and the preprocessed data.

The results of the K-nearest neighbor method are presented in Table 4, which gives the percentages of correctly classified compounds in each class of

TABLE 3

Feature reduction

Minimum no. of stationary liquid phases required for classification according to skeleton number	Minimum no. of stationary liquid phases required for classification according to functional groups
Zonyl E 7	Flexol 8N8
Flexol 8N8	Dow-Corning FS 1265 fluid
Tricresyl phosphate	Diethyleneglycol sebacate
Castorwax	Zonyl E 7
Sorbitol	
Diglycerol	
Dow-Corning FS 1265 fluid	
Polyphenyl ether — 6 Rings	
Diethyleneglycol sebacate	
XF-1150	

TABLE 4

Classification for functional groups by the K-nearest neighbor method

Functional group	Number of compounds	Percentages correctly classified									
		Original data					Modified data				
		1-NN	3-NN	4-NN	5-NN	10-NN	1-NN	3-NN	4-NN	5-NN	10-NN
Q	55	98	96	96	96	98	98	98	98	98	98
V,VH	38	66	58	53	53	26	84	87	84	84	84
VO	60	85	80	83	87	88	98	98	97	97	97
O	25	68	64	60	40	20	80	80	76	80	72
R	9	67	67	67	67	11	89	89	100	100	100
Alkanes and haloalkanes	12	33	25	25	17	17	75	67	67	67	67
All compounds	199	79	74	74	72	63	92	92	91	91	90

functional groups. It can be seen that classification with the original data set is good for alcohols and tolerable for esters, but unsatisfactory for all other types of compound. With the modified data set, however, good classifications are achieved for alcohols, esters and aromatic hydrocarbons, and satisfactory classifications for ethers and the combined class of aldehydes and ketones. Only the classification of the combined class of n-alkanes and haloalkanes is relatively poor; this is caused by the inhomogeneity of this group, which actually consists of two classes. Because of the small number of compounds of these types in the data library, the two classes had to be combined.

Next, the method of hierarchal clustering was applied. As an example, a cluster formed from the original data set is shown in Fig. 2a. It consists mostly of esters, but includes also three ketones, one aldehyde and one ether. The esters are C7 and C8 compounds, the other types are C6 and C7 compounds. An equivalent cluster formed from the modified data set is

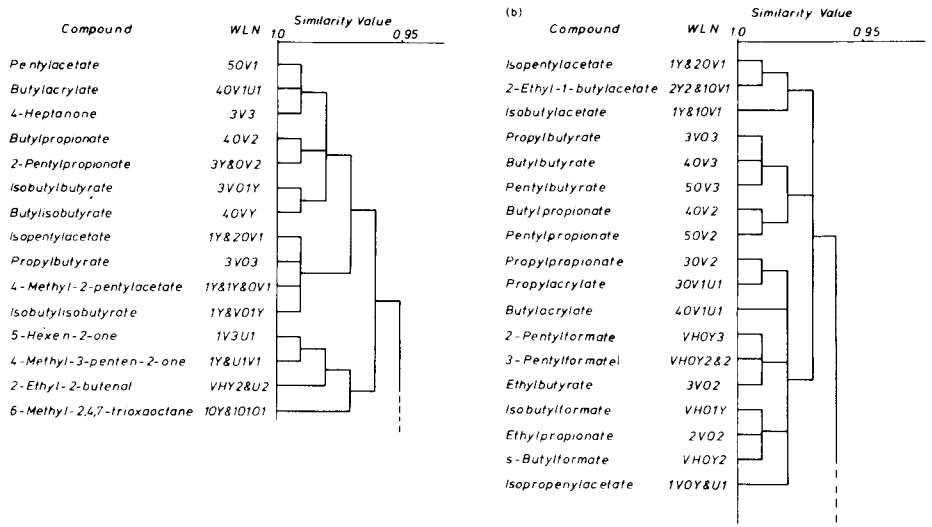


Fig. 2. Branching diagrams from the classification with the hierarchal clustering method: (a) original data (I_{Ri}); (b) pre-processed data ($\Delta_s I_{Ri}$).

shown in Fig. 2b. Both clusters have 4 compounds in common: butyl acrylate (4OV1U1), isopentyl acetate (1Y & 2OV1), butyl propionate (4OV2) and propyl butyrate (3VO3). The cluster of the modified data set consists entirely of esters, which have 5—9 carbon atoms. Figure 3 shows another corresponding cluster pair. Here again four compounds appear in both clusters: 2-heptanol (QY5), 2,2-dimethyl-1-pentanol (Q1X3), 3-heptanol (QY4 & 2) and 4-heptanol (QY3 & 3). The cluster from the original data set consists of isomeric hexanols and heptanols. The cluster of the modified data set

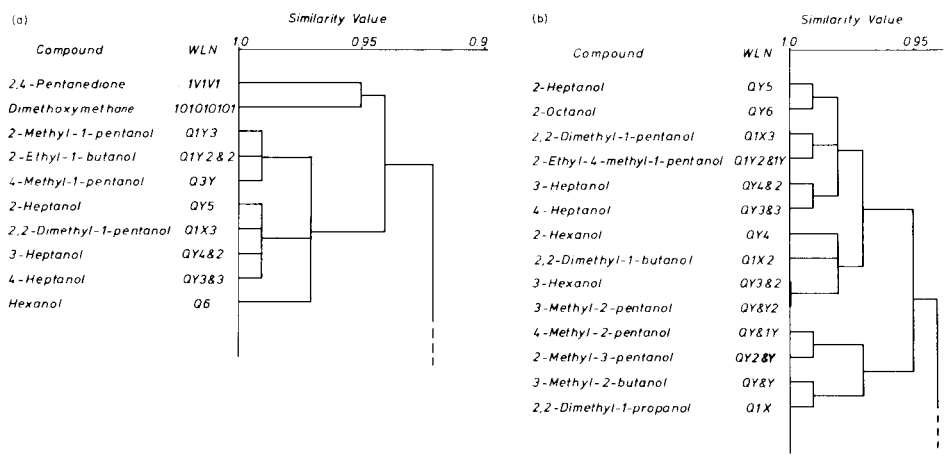


Fig. 3. Branching diagrams from the classification with the hierarchal clustering method: (a) original data (I_{Ri}); (b) pre-processed data ($\Delta_s I_{Ri}$).

consists of branched pentanols, hexanols, heptanols and octanols. It can be seen that in both cluster pairs the clustering in the original data set is more influenced by the chain length than by the type of functional group.

Similar results were found with another unsupervised learning method, the minimum spanning tree method. A typical example of the results obtained by this method for both data sets is presented in Table 5. Five compounds occur in both clusters: 1-butanol, 3-buten-1-ol, 1-propanol, 2-propen-1-ol and 2-buten-1-ol. The cluster in the original data set consists of isomeric butanols and pentanols; both n-alcohols and branched alcohols occur. The cluster in the modified data set consists of n-alcohols and unsaturated alcohols from methanol to octanol.

In the classification with the binary linear learning machine method, a significant parameter is the pattern/feature ratio. The result of classification will have no correlation with the trained property, if this ratio is too low (<3). With the relatively small data set, therefore, only the possibility of complete separation could be sought; the prediction probability could not be calculated. The results of these investigations are given in Table 6. The possibility of separation is characterized by the number of iteration cycles necessary to achieve complete (100%) separation. To eliminate the influence of the initial hyperplane on the number of iteration cycles, as for the determination of the skeleton number, two classifications were determined, the initial hyperplanes again being selected orthogonally. The average values

TABLE 5

Classification for functional groups with the minimum spanning tree method

Original data (I_R)		Modified data ($\Delta_s I_R$)	
Compound	WLN	Compound	WLN
3-Pentanol	QY2&2	Methanol	Q1
2-Pentanol	QY3	2-Propen-1-ol	Q2U1
3-Methyl-2-butanol	QY&Y	2-Methyl-2-propen-1-ol	Q1Y&U1
1-Penten-3-ol	QY2&1U1	3-Buten-1-ol	Q3U1
2,2-Dimethyl-1-propanol	Q1X	2-Buten-1-ol	Q2U2
1-Penten-4-ol	QY&2U1	Butanol	Q4
Isobutanol	QY4	Propanol	Q3
Butanol	Q4	3-Penten-1-ol	Q3U2
3-Buten-2-ol	QY&1U1	Pentanol	Q5
2-Methyl-3-butyn-2-ol	QX&&1UU1	Hexanol	Q6
3-Buten-1-ol	Q3U1	Heptanol	Q7
2-Butanol	Q1Y	Octanol	Q8
Propanol	Q3	2-Propyn-1-ol	Q2UU1
2-Methyl-2-propen-1-ol	Q1Y&U1		
2-Methyl-3-buten-2-ol	QX&&1U1		
2-Propen-1-ol	Q2U1		
2-Buten-1-ol	Q2U2		
2-Methyl-2-butanol	QX2		

TABLE 6

Classification of functional groups with the linear learning machine method

Functional group	Iteration cycles required for complete separation of two groups				
	Q	V,VH	VO	O	R
V,VH	55				
VO	23	366			
O	47	97	286		
R	31	23	25	28	
Alkanes and haloalkanes	35	9	6	6	15

from the two classifications are given in Table 6. It can be seen that the separation for functional groups is readily achieved.

Comparison of the results of the four pattern recognition methods in the classification of compounds for functional groups from gas chromatographic retention data shows that there is a difference between the methods based on a distance measure as the decision criterion (i.e. K-nearest neighbor, hierarchal clustering and minimum-spanning tree) and the linear learning machine method. The first three methods give similar results, with 80—90% correct classification, and a good impression of the spatial distribution of the samples. The linear learning machine method turns out to have superior classification properties, if two conditions are observed. The first condition is that the groups of points must be separable; the second is that the ratio between the number of points and the number of dimensions must be greater than 3 [35—38].

REFERENCES

1 P. C. Jurs and T. L. Isenhour, Chemical Applications of Pattern Recognition, Wiley—Interscience, New York, NY, 1975.
2 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 690.
3 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.
4 C. F. Bender and B. R. Kowalski, Anal. Chem., 45 (1973) 590.
5 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
6 G. S. Zander, A. J. Stuper and P. C. Jurs, Anal. Chem., 47 (1975) 1085.
7 J. R. McGill and B. R. Kowalski, Appl. Spectrosc., 31 (1977) 87.
8 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 42 (1970) 1387.
9 P. C. Jurs, Anal. Chem., 42 (1970) 1633.
10 J. B. Justice and T. L. Isenhour, Anal. Chem., 46 (1974) 223.
11 S. R. Heller, C. L. Chang and K. C. Chu, Anal. Chem., 46 (1974) 951.
12 P. C. Jurs, B. R. Kowalski and T. L. Isenhour, Anal. Chem., 41 (1969) 21.
13 H. Abe and P. C. Jurs, Anal. Chem., 47 (1975) 1829.
14 T. J. Stonham, I. Aleksander, M. Camp, W. T. Pike and M. A. Shaw, Anal. Chem., 47 (1975) 1817.
15 G. S. Zander and P. C. Jurs, Anal. Chem., 47 (1975) 1562.

16 T. F. Lam, C. L. Wilkins, T. R. Brunner, L. J. Soltzberg and S. L. Kaberline, Anal. Chem., 48 (1976) 1768.
17 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 42 (1970) 1387.
18 B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 1945.
19 D. R. Preuss and P. C. Jurs, Anal. Chem., 46 (1974) 520.
20 R. W. Lidell III and P. C. Jurs, Anal. Chem., 46 (1974) 2126.
21 H. B. Woodruff, S. R. Lowry, G. L. Ritter and T. L. Isenhour, Anal. Chem., 47 (1975) 2027.
22 J. S. Mattson, C. S. Mattson, M. J. Spencer and F. W. Spencer, Anal. Chem., 49 (1977) 500.
23 J. J. Leary, J. B. Justice, S. Tsuge, S. R. Lowry and T. L. Isenhour, J. Chromatogr. Sci., 11 (1973) 201.
24 S. R. Lowry, G. L. Ritter, H. B. Woodruff and T. L. Isenhour, J. Chromatogr. Sci., 14 (1976) 126.
25 D. L. Massart, P. Lenders and M. Lauwereys, J. Chromatogr. Sci., 12 (1974) 617.
26 S. Wold, J. Chromatogr. Sci., 13 (1975) 525.
27 H. A. Clark and P. C. Jurs, Anal. Chem., 47 (1975) 374.
28 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, 'ARTHUR', available from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry BG-10, Univ. Washington, Seattle, WA 98195 U.S.A., 1975.
29 W. O. McReynolds, Gas Chromatographic Retention Data, Preston Technical Abstracts, Evanston, IL, 1966.
30 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.
31 N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
32 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 1949.
33 W. L. Felty and P. C. Jurs, Anal. Chem., 45 (1973) 885.
34 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem. 42 (1970) 1387.
35 N. A. B. Gray, Anal. Chem., 48 (1976) 2265.
36 C. P. Weisel and J. L. Fasching, Anal. Chem., 49 (1977) 2114.
37 G. L. Ritter and H. B. Woodruff, Anal. Chem., 49 (1977) 2116.
38 C. F. Bender, H. D. Shepherd and B. R. Kowalski, Anal. Chem., 45 (1973) 617.
39 J. F. K. Huber and G. Reich, 14th Int. Symp. Advances in Chromatography, September 1979, Lausanne, Switzerland.

Analytica Chimica Acta, 122 (1980) 151—169
Computer Techniques and Optimization

COMPUTER IMPLEMENTATION OF SIMULATION MODELS FOR NON-LINEAR, NON-IDEAL CHROMATOGRAPHY
Part 2. Numerical Experiments and Results [1]†

J. C. SMIT and H. C. SMIT*

Laboratory for Analytical Chemistry, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam (The Netherlands)

E. M. DE JAGER

Institute of Applied Mathematics, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam (The Netherlands)

(Received 15th November 1979)

SUMMARY

The more practical aspects of simulation of discrete mathematical models are described. Particular attention is given to the model describing the behaviour of mass transport in chromatographic columns with a non-linear isotherm and with or without longitudinal dispersion. Rules for building general usable simulation software are explained. An outline of the program, without extensive treatment of program listings and flow diagrams, is given and experimental conditions and results are discussed.

Apart from mathematical considerations, the building of a simulation package, or the general development of software, has to satisfy several conditions. Firstly, and most importantly the package should be interchangeable, which means that a properly defined computer language has to be chosen without using machine-dependent tricks. Further, it is necessary that the software be capable of implementation without a lot of work either on a mainframe computer or on a minicomputer. Therefore, the software should preferably be written in modules, with the added advantage of simple testing procedures in the development phase.

In the present package, written in FORTRAN IV, three levels of module are distinguished. The first-level modules are the cosmetics of the software package, often written in an assembler language, which are used for printing date, time and text headings and also for options concerning application of sense switches. Consistent omission of these subroutines should not disturb the use of the rest of the program. The second-level modules are subroutines concerning special functions, which are used in the present program and not

---

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

defined in a standard FORTRAN library; these are applied for building the third-level modules. Subroutines of this kind are in most cases useful not only for this single application but also for general purposes. In the latter case they may be added to an object module library, though this results in higher demands on the commentary for the software. As software development usually involves teamwork, it is desirable to have a standard commentary in the headlines containing the purpose of the subroutine, directions for use, description of the parameters and, last but not least, the name of the responsible programmer, the last update, changes and literature references. The third-level modules are the subroutines or subprograms resulting in a special-purpose program. In the simulation package as described in this paper, there are eight main parts. Each part is written as a FORTRAN subroutine, but with the use of a common block structure each subroutine can be used as an overlay. Where meaningful, the subroutines are constructed in such a way that they also can function as an independent program. Table 1 shows a list of subroutines used in the simulation package.

FIRST-LEVEL MODULES

Such subroutines do not need extensive treatment, because they are too machine-dependent. For the sake of completeness, however, a short description is needed. The first in the series is a program part named HEADER for printing texts with capitals. The parts named DATE and TIMEP are based on assembler macro-routines for printing date and time, while RUNTEL has the function of an internal book-keeper. The last routine, named SWITS, can be used for suppressing diagnostics or activating jumps by means of a sense switch on the operator console.

SECOND-LEVEL MODULES

These modules are in general the most important elements for building a complete software package. In fact, the choice of these modules depends on the defined structure of the package. Nevertheless, a very important objective is that some subroutines should also be useful in other computer programs developed for quite different purposes. The best examples are the plot routines and some such mathematical routines are concerned with interpolation, differentiation, etc. In this paper, a brief description of the subroutines is presented with basic ideas concerning software. An extensive instruction manual with program texts is in preparation and will be available from the authors [2].

Mathematical routines

The first and most important routine is named RUNGE. This routine is the heart of the software package and gets its name from the time discretization used, a fourth-order Runge—Kutta method. Given the mathematical

TABLE 1

Subroutines in the simulation package

First-level modules	Second-level modules	Third-level modules
TIMEP	Mathematical routines	Main routines usable as overlay
DATE	SPLINE	CTREAD
HEADER	CURV	INJECT
SWITS	RUNGE	ISOEXP
RUNTEL	DIFF	ISOPLO
	Elution peak profile routines	NONLIN
	SMOM	RUN
	KUCERA	RISOT
	NONL	PLOTEL
	Isotherm routines	
	LANGMU	
	FREUND	
	SIGMO	
	Peak profile routines	
	AGAUSS	
	PPOISS	
	Plot routine (line printer plots)	
	CURVLU	
	Run time routines	
	TIMER	
	LIMITS	
	SHIFT	
	DECRE	
	ELUTE	

considerations outlined earlier [1], the main requirements are as follows. First, it is necessary to satisfy the divergence form

$$\frac{\partial C^t(t,z)}{\partial t} = -\frac{\partial V^* C^t(t,z)}{\partial z} + \frac{\partial^2 D^* C^t(t,z)}{\partial z^2} \tag{1}$$

with $V^* = \langle v \rangle \Big/ \left[1 + \frac{1-\epsilon}{\epsilon} \frac{C^s}{C^m}\right]$ and $D^* = D \Big/ \left[1 + \frac{1-\epsilon}{\epsilon} \frac{C^s}{C^m}\right]$

Here, $C^t(t,z)$ denotes the total concentration of the sample in the cross-section, $C^m(t,z)$ the sample concentration in the mobile phase, $C^s(t,z)$ the sample concentration in the stationary phase, and ϵ the cross-sectional area occupied by the fluid stream, while the functional dependence $C^s = f(C^m)$ is time- and place-invariant. The coefficients $\langle v \rangle$ and D are independent of the concentration, denoting the average fluid velocity over the corresponding cross-sectional area and the dispersion effect (a measure of the axial molecular diffusion and the eddy mixing effect), respectively. The final coefficient is assumed to be independent of the fluid velocity $\langle v \rangle$. A fourth-order Runge—Kutta discretization of the time-dependent part has to be used,

together with a first-order Newton—Backward and a second-order Stirling discretization for the velocity part and the dispersion part, respectively. Finally, the introduction of moving coordinates for minimizing the numerical dispersion effect is inevitable. To keep the divergence form, the use of an expression in terms of total concentration C^t is the most convenient. From the chromatographic point of view, however, the main interest lies in the elution profiles in terms of mobile phase concentrations C^m and so it is necessary to calculate the elution profiles of C^m from those of C^t. For this purpose the routines LANGMU, FREUND, SIGMO and CURV were developed and will be discussed later. For efficient calculations, seperate algorithms for the linear and the non-linear case are inserted into the subroutine, while in the non-linear case the Runge—Kutta method is modified in some respects.

Carrying out the suggested discretization of eqn. (1) yields the next algorithm (see also Part 1 eqns. 33 and 34).

$$y_n^{k+1} = y_n^k + h\ (1/6)\ \{K_0^k\ (n) + 2K_1^k\ (n) + 2K_2^k\ (n) + K_3^k\ (n)\}$$

$$\text{with } K_0^k\,(n) = \{y_n^k\}*\{w_n\},\ K_1^k\,(n) = \{y_n^k + (1/2)h\ K_0^k\,(n)\}*\{w_n\}$$

$$K_2^k\,(n) = \{y_n^k + (1/2)h\ K_1^k(n)\}*\{w_n\},\ K_3^k\,(n) = \{y_n^k + h\ K_2^k\,(n)\}*\{w_n\} \qquad (2)$$

where y_n^{k+1} denotes $C^t((k+1)\Delta t,z)$, y_n^k denotes $C^t(k\Delta t,z)$ and h the time step, Δt. When the Runge—Kutta method is used in a straightforward way, the estimated derivatives $K_0^k(n)$, $K_1^k(n)$, $K_2^n(n)$ and $K_3^n(n)$ are obtained from the convolution of weighting function w_n with

$$y_n^k := [y_n^k]_0\ ;\ y_n^k + h\ K_0^k(n)/2 := [y_n^{k+1/2}]_1$$

$$y_n^k + h\ K_1^k(n)/2 := [y_n^{k+1/2}]_2\ ;\ y_n^k + h\ K_2^k(n) := [y_n^{k+1}]_3$$

where the suffixes 0—3 denote, respectively, the value at time t, the first estimated value at $t = t + \frac{1}{2}\Delta t$, the second estimated value at $t = t + \frac{1}{2}\Delta t$ and the estimated value at $t = t + \Delta t$. In the linear case the weighting function is independent of the concentration, and

$$w_n = \begin{cases} 0 & n < -1 \\ \alpha = \dfrac{D^*}{l^2} & n = -1 \\ \beta = -\dfrac{V^*}{l} - \dfrac{2D^*}{l^2} & n = 0 \\ \gamma = \dfrac{V^*}{l} + \dfrac{D^*}{l^2} & n = +1 \\ 0 & n > +1 \end{cases} \qquad (3)$$

where D^* and V^* are constants. In non-linear cases, however, a better estimation should be made by the convolution of $[y_n^k]_0$, $[y_n^{k+1/2}]_1$, $[y_n^{k+1/2}]_2$ and $[y_n^{k+1}]_3$ with a weighting function, dependent on the intermediate estimated values. Thus in the non-linear case the values K_0^k, K_1^k, K_2^k, and K_3^k are obtained

from

$$K_0^k = [y_n^k]_0 * \{w_n\}_0 \;;\; K_1^k = [y_n^{k+1/2}]_1 * \{w_n\}_1 \;;\; K_2^k = [y_n^{k+1/2}]_2 * \{w_n\}_2$$
$$K_3^k = [y_n^{k+1}]_3 * \{w_n\}_3 \tag{4}$$

where $\{w_n\}_0$, $\{w_n\}_1$, $\{w_n\}_2$, and $\{w_n\}_3$ are now weighting functions, dependent on $[y_n^k]_0$, $[y_n^{k+1/2}]_1$, $[y_n^{k+1/2}]_2$, $[y_n^{k+1}]_3$, respectively.

In the linear case, where the weighting function is constant, the convolutions of eqn. (2) can be carried out very easily and after some algebraic calculations a nine-point explicit difference scheme is obtained

$$y_n^{k+1} = \{y_n^k\} * \{v_n\} \tag{5}$$

where v_n is a new weighting function which can be expressed in terms of the weighting function w_n; after some calculations the ultimate result is

$$v_n = \begin{cases} 0 & n < -4 \\ \alpha^4/24 & n = -4 \\ \alpha^3(1+\beta)/6 & n = -3 \\ \alpha^2(1+\alpha\cdot\gamma/3+\beta(1+\beta/2))/2 & n = -2 \\ \alpha(1+\alpha\cdot\gamma/2+\beta(1+\alpha\cdot\gamma/2+\beta(0.5+\beta/6))) & n = -1 \\ \alpha\cdot\gamma(1+\alpha\cdot\gamma/4+\beta+\beta^2/2)+\beta(1+\beta(0.5+\beta(1/6+\beta/24))) & n = 0 \\ \gamma\cdot(1+\alpha\cdot\gamma/2+\beta(1+\alpha\cdot\gamma/2+\beta(0.5+\beta/6))) & n = 1 \\ \gamma^2(1+\alpha\cdot\gamma/3+\beta(1+\beta/2))/2 & n = 2 \\ \gamma^3(1+\beta)/6 & n = 3 \\ \gamma^4/24 & n = 4 \\ 0 & n > 4 \end{cases} \tag{6}$$

The following mathematical routines are named SPLINE and CURV; both are interpolation routines. The first, which is particularly useful for the interpolation of experimental non-equidistant data points, is a conversion of an ALGOL program published by Reinsch [3]. Subroutine CURV is a third-order interpolation routine with the advantages of minimum oscillation, a continuous first derivative and simpler applicability in comparison with the SPLINE interpolation. The interpolated value for $x_j < x < x_{j+1}$ is evaluated by means of a third-degree polynomial $y = ax^3 + bx^2 + cx + d$. The coefficients a, b, c and d are calculated with the aid of the points $P_j(x_j, y_j)$, $P_{j+1}(x_{j+1}, y_{j+1})$ and the derivatives $y'(x_j)$ and $y'(x_{j+1})$ (see Fig. 1). The value for $y'(x_{j+1})$ is obtained by also using point $P_{j+2}(x_{j+2}, y_{j+2})$ and the Lagrange formula. In order to simplify the calculations, a scaling and shifting factor is introduced such that $x_j = 0$ and $x_{j+1} = 1$. By differentiating the Lagrange formula

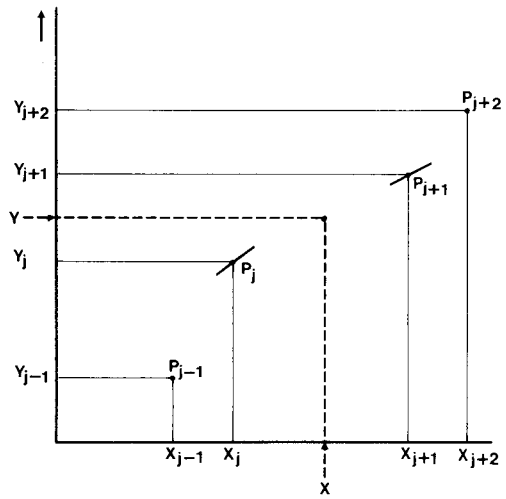


Fig. 1. Outline of the interpolation routine CURV.

$$y = \frac{(x - x_{j+1})(x - x_{j+2})}{(x_j - x_{j+1})(x_j - x_{j+2})} \cdot y_j + \frac{(x - x_j)(x - x_{j+2})}{(x_{j+1} - x_j)(x_{j+1} - x_{j+2})} \cdot y_{j+1}$$

$$+ \frac{(x - x_j)(x - x_{j+1})}{(x_{j+2} - x_j)(x_{j+2} - x_{j+1})} y_{j+2} \qquad (7)$$

$(j = 0, 1, 2, \ldots N)$
where N is the total number of data points. For $x_j = 0$ and $x = x_{j+1} = 1$

$$y'(1) = y_{j+1} - y_j + \frac{y_j}{x_{j+2}} + \frac{y_{j+1}}{(1 - x_{j+2})} - \frac{y_{j+2}}{x_{j+2}(x_{j+2} - 1)} \qquad (8)$$

With the coordinates of points P_j and P_{j+1} and the derivatives $y'(0)$ and $y'(1)$, the coefficients a, b, c and d are calculated, which gives

$d = y(0)$; $c = y'(0)$; $a = y'(1) + y'(0) + 2\{y(0) - y(1)\}$;
$b = y(1) - a - c - d$

The interpolated value has to be corrected only by the scaling and shifting factor. The interpolated values for $x_j < x < x_{j+1}$, where j is 0 or $n - 1$, are obtained by a second-degree polynomial.

The mathematical routine DIFF is a differentiation routine for equidistant data points and is based on Stirling's formula of sixth degree, i.e.

$$\mathrm{d}\, y_n/\mathrm{d}z = (\delta_0 - \delta_2/6 + \delta_4/38 - \delta_6/148)/2\Delta z \qquad (9)$$

where $\delta_0 = y_{n+1} - y_{n-1}$
$\delta_2 = y_{n+1} - 2y_n + y_{n-1}$
$\delta_4 = y_{n+2} - 4y_{n+1} + 6y_n - 4y_{n-1} + y_{n-2}$
$\delta_6 = y_{n+3} - 6y_{n+2} + 15y_{n+1} - 20y_n + 15y_{n-1} - 6y_{n-2} + y_{n-3}$

More detailed information is given by Scheid [4].

Elution peak profile routines

Peaks are usually characterized by their statistical moments, defined as

$$m_0 = \sum_{n=-\infty}^{+\infty} y_n \; ; \; m_j = \sum_{n=-\infty}^{+\infty} n^j y_n \; (j = 1,2,3,4, ...)$$

$$\tilde{\mu}_1 = m_1/m_0; \; \tilde{\mu}_2 = m_2 - m_1^2;$$

$$\tilde{\mu}_3 = m_3 - 3m_1m_2 + 2m_1^3 \; ;$$

$$\tilde{\mu}_4 = m_4 - 4m_1m_3 + 6m_1^2 m_2 - 3m_1^4$$

$$\text{Skewness} = \tilde{\mu}_3/(\tilde{\mu}_2^3)^{1/2} \; ; \; \text{Kurtosis} = \tilde{\mu}_4/(\tilde{\mu}_2)^2$$

The subroutine SMOM yields the values of these peak parameters. The subroutine KUCERA yields the parameters of the elution profiles, resulting from the exact solution of the linearized transport eqn. (1). This routine is based on the original work of Kučera [5] and the final formulae are given by

$$\mu_1 = a + b$$

$$\tilde{\mu}_2 = b(a + 2b) \; ; \; \tilde{\mu}_3 = b^2(3a + 8b) \; ; \; \tilde{\mu}_4 = 3b^2(a^2 + 9ab + 20b^2)$$

with $a = L/V^*$, and $b = 2D^*/(V^*)^2$, where L denotes the length of the column, V^* the migration velocity and D^* the effective dispersion, already defined in eqn. (1).

The subroutine NONL yields the concentration distributions C^t, C^m and C^s as well as a function of time t for several values of z, as well as a function of z for several values of t. The calculation is based on the theoretical result of Part 1 [1], where with the aid of the method of characteristics an explicit formula for C^t was given (compare Part 1, eqn. 100). C^s and C^m were correlated to each other by a non-linear Langmuir isotherm and physical dispersion was neglected.

Isotherm routines

In order to investigate the effects of non-linearity by means of a simulation package, the results stemming from a linear isotherm and a non-linear isotherm should be compared. As to the latter, a Langmuir isotherm, a Freundlich isotherm, a sigmoidal or an experimental isotherm can be selected. Beginning with the last in the series, the experimental isotherm data can be inserted into the program. After conversion, with the use of subroutine SPLINE, a 500-point search table is obtained for simulation purposes.

To generate theoretical isotherms correlating C^t and C^m, three subroutines are created: LANGMU for isotherms of the Langmuir type, FREUND for isotherms of the Freundlich type and SIGMO for sigmoidal isotherms. Normally, an isotherm is represented as $C^s = f(C^m)$; in the present simulation, however, the functional dependence $C^s/C^m = g(C^t)$ is preferable. From the point of view of numerical analysis, a short description of the algorithms leading to this kind of theoretical isotherm is worth mentioning. In relation to the input variables C^t, the isotherm parameters and the phase ratio ϵ, the mobile phase and stationary phase concentrations are calculated.

The subroutine LANGMU is based on the application of the Newton–Raphson iterative procedure for obtaining the value of C^m from the formula

$$C^t = \epsilon C^m + (1-\epsilon)C^s = \epsilon C^m + (1-\epsilon)\,AC^m/(1+BC^m) \tag{10}$$

where for C^s the Langmuir relation $C^s = A\,C^m/(1+BC^m)$ has been substituted. Putting $T_1 = B\epsilon$ and $T_2 = \epsilon + (1-\epsilon)A - BC^t$, one gets $T_1(C^m)^2 + T_2C^m - C^t = 0$ and the Newton–Raphson procedure yields

$$C^m_{new} = C^m_{old} - [(C^m_{old})^2\,T_1 + C^m_{old}\,T_2 - C^t]/[2T_1C^m_{old} + T_2] \tag{11}$$

as the start condition of the iterative procedure, which is decisive for the number of iterations. The best start condition is the highest of the following two start values

$$C^m_{old,1} = \frac{C^t}{\epsilon} - \frac{(1-\epsilon)}{\epsilon}\frac{A}{B} \text{ and } C^m_{old,2} = \frac{C^t}{\epsilon + (1-\epsilon)A}$$

These values are obtained by the assumptions $BC^m \gg 1$ and $BC^m \ll 1$, respectively. The iteration stops at a difference of $|10^{-10}|$ between C^m_{old} and C^m_{new}.

Subroutine FREUND also calculates the values of C^m and C^s for a given input variable C^t in relation to the isotherm parameters and the phase ratio ϵ, and is quite similar to the subroutine LANGMU. Instead of the Langmuir isotherm, the Freundlich isotherm, viz. $C^s = A(C^m)^B$ for $0 \leqslant B \leqslant 1$, is substituted for C^s. In this case it is appropriate to calculate $P = \ln C^m$ and an elementary derivation yields

$$\epsilon\,e^P + (1-\epsilon)A\,e^{PB} - C^t = 0 \tag{12}$$

and hence the following Newton–Raphson iteration scheme is obtained

$$P_{new} = P_{old} - \frac{T_3 + T_4 - T_2}{T_3 + B\,T_4} \tag{13}$$

with $T_1 = A(1-\epsilon)/\epsilon$; $T_2 = C^t/\epsilon$; $T_3 = e^{P_{old}}$; $T_4 = T_1 \cdot e^{(P_{old}\cdot B)}$

Rewriting eqn. (12) gives

$$\frac{C^t}{(1-\epsilon)A} = \frac{\epsilon}{(1-\epsilon)A} \cdot C^m + (C^m)^B$$

and so for

$$C^m \ll (C^m)^B \;,\; C^m = \exp\left\{\left[\ln \frac{C^t}{(1-\epsilon)A}\right]/B\right\} = S_1$$

$$C^m \gg (C^m)^B \;,\; C^m = C^t/\epsilon = S_2$$

The start condition is now chosen as $P_{old} = S_1\,S_2/(S_1 + S_2)$.

In the subroutine SIGMO we consider sigmoidal isotherms, which have the shape of a Langmuir isotherm for $0 \leqslant C^t \leqslant (1/2)\,C^t_{max}$ and its mirror image for $(1/2)\,C^t_{max} \leqslant C^t \leqslant C^t_{max}$. This sigmoidal isotherm can be constructed with the aid of a mirror image of a Langmuir isotherm.

Peak profile routines

Although the possibility of the introduction of several injection profiles is very useful, the details of only two subroutines are discussed here, namely AGAUSS and PPOISS. For the introduction of injection profiles, expressed more mathematically in terms of initial conditions, the subroutine INJECT, described in a later section should be consulted.

The first subroutine evaluates the integral of a Gaussian probability function between the integration limits z_n and $z_n + \Delta z$, with z_n equidistantly distributed over the length of the column. The input variables are the zero moment m_0, the first normalized moment μ_1, the standard deviation $(\tilde{\mu}_2)^{1/2}$ and the integration limits. The second subroutine evaluates a Poisson probability function in a similar way. These subroutines are modified versions of FORTRAN function subprograms of the same name published by Bevington [6].

Plot routines

For plotting purposes, a standard plot routine CURVLU written in FORTRAN IV was developed for making point plots on several kinds of logical units (line printer, video terminal, etc.). The plot software for the Calcomp plotter, used for publishing the present results is not a part of the simulation software and is only commercially available.

Run time routines

The last of the series of subroutines belonging to the second-level modules, are TIMER, LIMITS, SHIFT, DECRE and ELUTE. The subroutine TIMER yields a summation of the steps Δt and may give the possibility of extending the software package in order to investigate the effects of non-stationarity of the concentration distribution and to develop gradient elution techniques. Although this kind of simulation problem has not yet been tackled in this work, these possibilities should not be excluded. The subroutine can be coupled with time-dependent variables.

The next subroutine named LIMITS indicates crossing of the array space limits and activates the subroutine SHIFT for a shift operation, or DECRE for decreasing the array where the actual place profile has been stored. The subroutine ELUTE evaluates the elution profile $C^t(t,z)$ at a fixed boundary $z = L$.

THIRD-LEVEL MODULES

As mentioned above, this kind of module, also written in FORTRAN IV, may be used as a subroutine but also as an overlay. The first subroutine named CTREAD is a program part where all parameters necessary for carrying out the simulation are inserted and can be changed interactively. A block is filled with all parameters, e.g., column length, phase ratio, fluid velocity, dispersion, the kind of isotherm including the isotherm parameters,

TABLE 2

Parameters for subroutine CTREAD

Column parameters	Column length, phase ration
Mass transport parameters	Dispersion, velocity (flow)
Injection parameters	Rectangular, Gaussian, Poisson left-bounded, Poisson right-bounded, triangular, pulse, experimental
Isotherm parameters	Linear, Langmuir, Freundlich, sigmoidal, experimental
Simulation parameters	Time step DELTT, place step DELTZ, moving coordinates, array space definition, diagnostics

kind of injection profile, etc. In a dialogue with the program, the program presents suggestions for certain values of the parameters in order to guarantee stability of the calculations, and asks for printing instructions for extensive diagnostics during run time (see Table 2). The second module named INJECT generates the injection profile or initial condition in relation to the parameters in the common block (see Fig. 2). The profiles are stored in the working array and are protected against crossing the array limits. In this case, crossing does not activate the subroutine DECRE, but a new array space is defined automatically.

The use of rectangular injection profiles is appropriate for the investigation of the step response or the so-called break-through curves, of the Poisson right-bounded profiles for the investigation of subsequent delivery, and of the triangular profiles for the study of the influence of pseudo-random binary input sequences (PRBS) [7]. Furthermore, it is possible to insert any injection profile from the keyboard of the operator console.

The third module ISOEXP gives, by means of a SPLINE interpolation, a 500-point isotherm, which is obtained from a maximum of 50 experimental non-equidistant data pairs (C^s, C^m). This subroutine includes facilities for reading or writing converted data from or to disc files and also a facility for assigning experimental data to logical units.

The next module ISOPLO plots (see Fig. 3) the isotherm $C^s = f(C^m)$ and all information following from these isotherms, such as the functional dependence of the velocities $V^*(C^t)$, $V^{**}(C^t)$ and the derivative of C^s with respect to C^m. The introduced velocity $V^{**}(C^t)$ denotes a sort of migration velocity and is defined as

$$V^{**}(C^t) = \langle v \rangle \Big/ \left[1 + \left(\frac{1-\epsilon}{\epsilon}\right)\frac{dC^s}{dC^m}\right]$$

with only a mathematical meaning [1].

A very rough stability criterion for the non-linear case may be calculated with the aid of the following ideas. In order to obtain convergence of the numerical results it was necessary to choose at the point P $(n\Delta z, k\Delta t)$ the differences Δz and Δt according to $(\Delta z/\Delta t) \leqslant V^{**}_{max}$, where $1/V^{**}_{max}$ denotes

the slopes of the characteristic for the concentration at the point P [1, 8]. In order to prevent oscillations in the neighbourhood of a discontinuity the algorithm must never lead to values of y_n^{k+1} which are negative or which are larger than y_n^k. This leads to the following conditions, which are easily derived from the algorithm in the case of a pulse as initial condition (cf. eqn. 56 [1])

$$V_{ref} \leqslant V^*_{max} + (2D^*_{max}/\Delta z) \tag{14}$$

$$V_{ref} \geqslant V^*_{max} + (2D^*_{max}/\Delta z) - (\Delta z/\Delta t) \tag{15}$$

where V_{ref} refers to the translational velocity of the moving coordinate system. Taking for V_{ref} the lowest migration velocity V^*_{min}, condition (14) is fulfilled. Combining eqn. (15) with $\Delta z/\Delta t \leqslant V^{**}_{max}$ and $V_{ref} = V^*_{min}$ gives

$$V^{**}_{max} \geqslant \Delta z/\Delta t \geqslant V^*_{max} + (2D^*_{max}/\Delta z) - V^*_{min} \tag{16}$$

and hence

$$\Delta z \geqslant 2D^*_{max}/(V^{**}_{max} - V^*_{max} + V^*_{min}) \tag{17}$$

and

$$\Delta t \leqslant \Delta z/[V^*_{max} - V^*_{min} + (2D^*_{max}/\Delta z)] \tag{18}$$

The criteria (17) and (18) are used for preventing possible occurrence of instabilities in the calculations.

The module RUN is a combination of RUNGE, LIMITS, DECRE, SHIFT, TIMER and ELUTE, where the last subroutine gives the elution profile (see Fig. 4). The elution profile obtained in the program part RUN is stored in the common block. At first, by calling subroutine or overlay PLOTEL, the elution profile is plotted and is also written in an assigned logic unit (disc file or tape punch). The statistical moments are also calculated and printed. Now the elution profile can be compared with, for instance, the computed exact solution, created by calling NONLIN. Another possibility is the recalculation of the assumed isotherm from the peak shape, as suggested by Huber and Gerritse [9], with the aid of subroutine RISOT. An extra advantage of the subroutine RISOT is that it is also useful for calculating isotherm non-linearities from experimental chromatographic data [10].

NUMERICAL EXPERIMENTS (SIMULATION) AND RESULTS

Short description of the hardware

The development of the simulation package described, which is named SAM, was carried out on a minicomputer system (Varian Data Machines) equipped with a core memory of 96K words of 16 bits, two discs with a total capacity of 2.4M words and Vortex-E as operating system. Peripherals used in the configuration are a Tally line-printer, several video terminals, a tape reader, a card reader, a tape punch and a microprocessor-based data

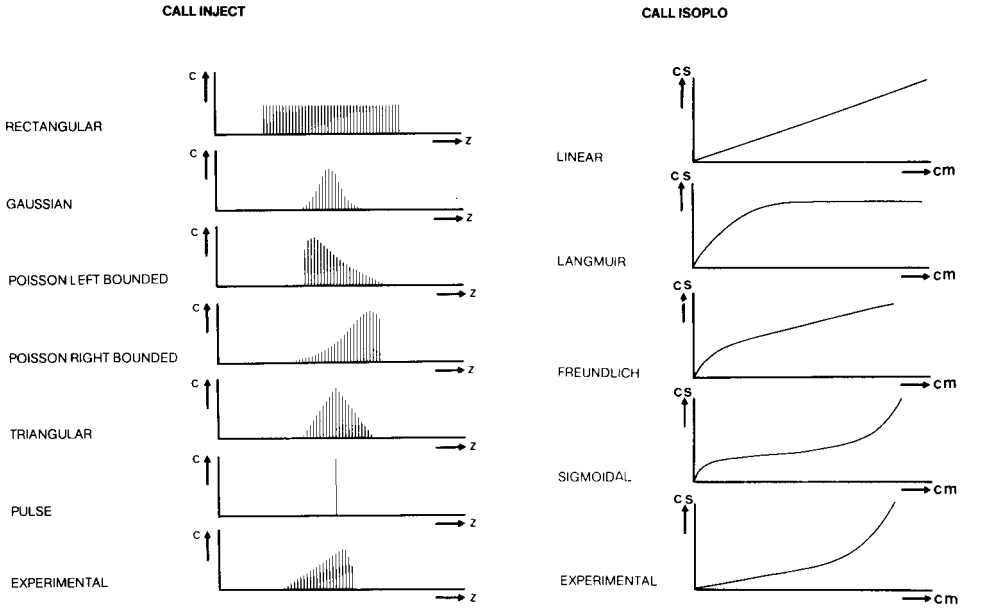


Fig. 2. Parameters of the common block for module INJECT. The injection parameters give the zero moment, first normalized moment and second central moment.

Fig. 3. Isotherm plots.

acquisition system. The software, with respect to input/output activities, is realised by the use of FORTRAN unit numbers in the read/write statements. These FORTRAN unit numbers have to be defined separately in the first lines of each program part, because these device numbers, or logic unit numbers, are machine-dependent. The results of the numerical experiments presented in this paper were plotted by means of the Calcomp plot device available at the SARA Computer Centre (Amsterdam).

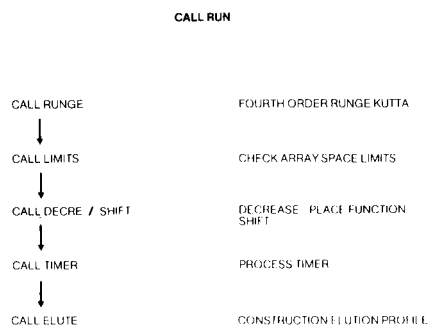


Fig. 4. The RUN module.

Moving coordinate systems

The first experiment to be described is the simulation of a column with a linear distribution isotherm which has a slope $A = 1$ and a phase ratio $\epsilon = 0.5$. Physical dispersion is neglected and so a pure translation of the initial condition given by the injection profile should be obtained. In this case a pulse of width $\Delta z = 1$ and zero statistical moment $m_0 = 10$ is injected. Figure 5 shows the concentration profiles at $t = 4, 8, 12, 16, \ldots\ldots 40$ s, resulting from the application of subroutines CTREAD, INJECT and RUN. The algorithm was realised by the subroutine RUN without using the mathematical device of moving coordinates and so an inadmissible numerical dispersion effect is observed.

It is also possible to compute the concentration profiles with the aid of the so-called chamber model or ideal mixer model. This model is given by the formula

$$C^t(t,n) = C_0^t\,[(t/\tau)^{n-1}/(n-1)!]\cdot e^{-t/\tau} \qquad (19)$$

where $C^t(t,n)$ is the total sample concentration at time t in mixer n; C_0^t is the initial concentration in the first mixer ($C^t(t,n) = 0$ for $n = 2,3,4, \ldots$); τ is the

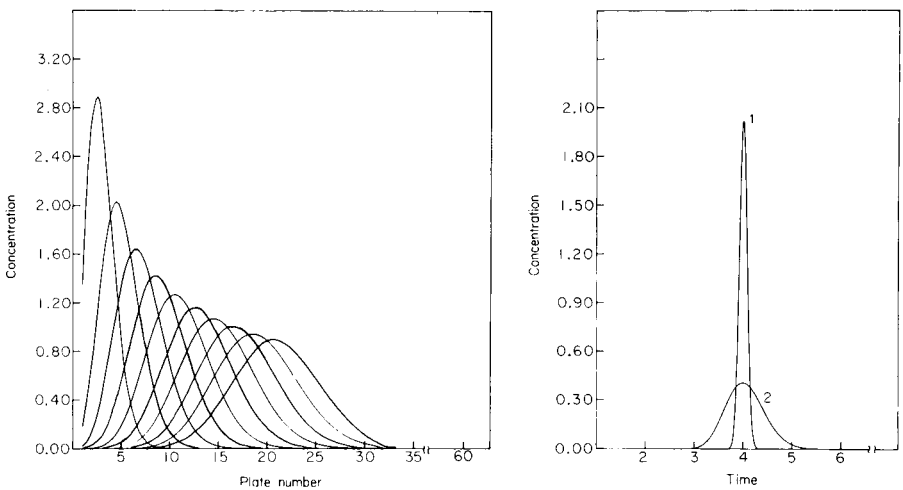


Fig. 5. Place functions calculated by the simulation program SAM for $t = 4, 8, 12, \ldots\ldots 40$ s; $\langle v \rangle = 1$ cm s^{-1}; $\epsilon = 0.5$; linear isotherm with slope $A = 1$; injection profile pulse-shaped with $m_0 = 10$; $\Delta t = 0.5$ s; $\Delta z = 1$ cm; without moving coordinates.

Fig. 6. Two elution profiles calculated by the simulation program SAM. Peak 1 is with moving coordinates and peak 2 without moving coordinates; dispersion = 0.1; $\langle v \rangle = 5$ cm s^{-1}; $\epsilon = 0.5$; $L = 100$ cm; linear isotherm with slope $A = 1$; injection profile pulse-shaped with $m_0 = 10$; $\Delta t = 0.25$ s; $\Delta z = 0.625$ cm.

mixer time constant $\Delta z/V_c$; n is the rank number of the mixers (n = 1,2,3,4, ... N); N is the total number of mixers; Δz is the mixer width and V_c is the migration velocity, assuming a linear distribution isotherm.

Comparison of the numerical values of $C^t(t,n)$ obtained from eqn. (19) with the results of Fig. 5 reveals that the algorithm carried out by subroutine RUN yields a perfect representation of the ideal mixer model; the deviation is within the calculation accuracy of the algorithm.

Figure 6 presents the elution profiles evaluated with and without the device of a moving coordinate system; Table 3 gives numerical values of the statistical moments of both profiles and includes the results obtained from Kučera's relationships (see above) programmed in subroutine KUCERA. The observation that the results for the elution profile 1, apart from those of the skewness, are in very good agreement with the Kučera results shows that the introduction of moving coordinates is necessary for the prevention of numerical dispersion. The cause of the disagreement in the results for the skewness lies in the fact that the elution profiles were unfortunately truncated in an asymmetric way. It can therefore be concluded that the numerical dispersion effect is reduced to zero by the use of moving coordinates in linear cases. The numerical dispersion effect may also be reduced by taking a larger number of mixers. This results in higher and smaller elution profiles, but also, unfortunately, in a proportional increase of computing time. The introduction of the condition [1] $V_c = \Delta z/\Delta t$, where V_c is the migration velocity in the linear case, also decreases the numerical dispersion effect but yields instabilities and negative concentrations. In the non-linear case, the reducing effect of the use of moving coordinates depends on the curvature of the isotherm. The larger the curvature of the isotherm, the more numerical dispersion has to be considered. In practice, a stable algorithm with minimal numerical dispersion is preferable to an unstable one. However, numerical dispersion cannot always be avoided, for example, in the case when a Freundlich isotherm is used in the simulation of non-linear mass transport. The slope of the isotherm is then infinite at zero concentration and so the migration velocity V^*_{min} becomes zero; because the velocity of the moving coordinate system is equal to V^*_{min}, numerical dispersion occurs.

TABLE 3

Numerical values of the statistical moments of both profiles

Moments	Moving coordinates elution profile 1	Non-moving coordinates elution profile 2	Theoretical
m_0	9.9997	9.9798	10.0000
μ_1	40.0160	40.5236	40.0160
$\tilde{\mu}_2$	0.6405	15.6758	0.6405
Skewness	1.6049	472.8380	0.0599
Kurtosis	3.1039	3.1064	3.0060

Non-linear isotherms

In this section, seven numerical experiments are discussed concerning non-linear mass transport, where, however, physical dispersion has been neglected. The model used involves a strongly curved Langmuir isotherm with parameters $A = 100$ and $B = 10$. Seven elution profiles evolving from seven pulse injection concentrations, viz. 32, 16, 8, 4, 2, 1 and 0.5 mol/unit volume, are shown in Fig. 7. These plots were obtained by calling the subroutines CTREAD, INJECT, ISOPLO, RUN and PLOTEL. Despite the use of moving coordinates there is still numerical dispersion as appears from the steepness of the peak, most clearly apparent for the injections with lower concentrations. The zero statistical moments (see Table 4) of these elution profiles do not differ very much from those of the injection profiles, and so the tails are probably distorted a little by numerical dispersion. The reconstruction of the isotherm from the peak shapes by calling the subroutine RISOT shows that the results are quite acceptable.

The elution profiles presented in Fig. 7 and Table 4 can be compared with the elution profiles to be obtained from the exact solution given in Part 1 eqn. (100), which is programmed in subroutine NONL. By calling subroutine or overlay NONLIN, which is a combination of NONL, CURVLU and SMOM, the elution profile including numerical values of statistical moments is obtained. The comparison immediately gives the contribution of the numerical dispersion to the shape of the elution profile.

Wherever physical dispersion is not neglected in these models, the numerical dispersion is not very important, because in practice the former has much more influence than the latter. If the influence of the numerical dispersion on the shape of the elution profile is known, then in principle this distorting effect can be accounted for by decreasing the physical dispersion with the aid of a certain factor smaller than 1. For practical chemical purposes this correction is hardly important, but from the mathematical point of view it might be interesting.

TABLE 4

Statistical moments referring to the experimental results presented in Fig. 7; moments are calculated from the mobile phase concentrations $C^m(t)$ except for m_0 which is the zeroth moment of the flux $C^m(t)v$

Elution profiles	1	2	3	4	5	6	7
m_0	31.7323	15.9883	7.9948	3.9981	2.0015	1.0011	0.5010
μ_1	40.6767	58.5843	74.6863	87.8505	97.9547	105.4120	110.6990
$\tilde{\mu}_2$	486.4230	407.1500	265.2840	155.8980	86.4608	46.6132	25.2635
Skewness	2.8479	2.2478	1.4002	0.9384	0.6170	0.4680	0.3301
Kurtosis	3.4053	2.9608	2.6802	2.5240	2.4181	2.3342	2.2666
Relative error range reconstructed data points with RISOT (%)	5.3—8.7	0—5.5	0—5.0	0—4.9	0—4.7	0—4.4	0—4.0

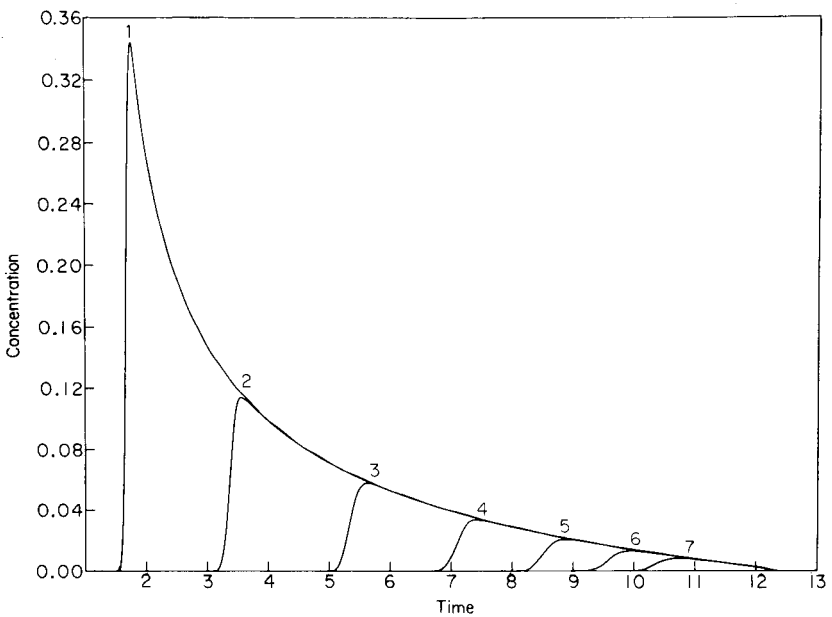


Fig. 7. Seven elution profiles calculated by the simulation program SAM with an increasing input concentration; no dispersion; $\langle v \rangle = 5$ cm s^{-1}; $\epsilon = 0.9$; $L = 50$ cm; Langmuir isotherm with $A = 100$ and $B = 10$; pulse-shaped injection profile with m_0 values of 32, 16, 8, 4, 2, 1, and 0.5 mol/volume unit for curves 1—7, respectively; $\Delta t = 0.1$ s; $\Delta z = 1.0$ cm; with moving coordinates.

The non-linear case with physical dispersion

In this section, the results of 17 numerical experiments are described in order to investigate the shape of the elution profiles in relation to the curvature of Langmuir isotherms. The influence of physical dispersion is taken into account. Figure 8 shows a family of isotherms described by $C^s = AC^m/(1 + BC^m)$, where $A = 1$ and the values of B, determining the curvature, vary from 0 to 9.9. The results reproduced in Fig. 9 and in Table 5 give a good impression of the possibilities of the simulation package. The numbers in the first column of Table 5 refer to the peaks in Fig. 9, where the peak numbering is done from right to left. The second column contains the numerical values of isotherm parameter B. All moments, except the zeroth moment m_0, are calculated values based on the mobile phase concentrations $C^m(t)$. The values of m_0 are calculated with the aid of the mass flux $\langle v \rangle C^m(t)$ at the boundary $z = L$. These values of m_0 are directly comparable with the values of m_0 of the injection profiles. The unpredictable fluctuations in the skewness (column 6, Table 5) are caused by the arbitrary choices of the truncation of the elution profiles. As the skewness, characterizing the asymmetry of the profiles, is very sensitive to asymmetrical truncation, proper truncation limits require careful attention. The first row in the table contains the statistical moments obtained by subroutine KUCERA for the linear case ($B = 0$).

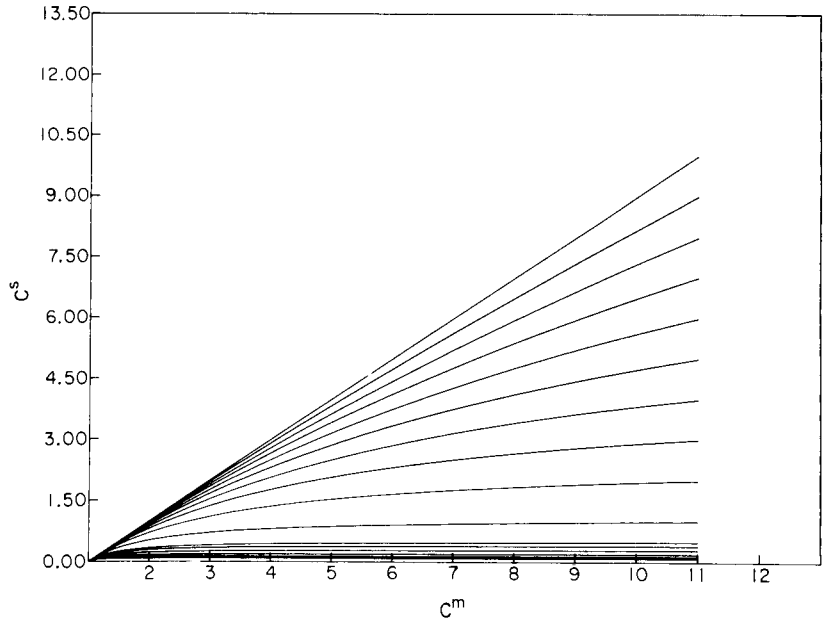


Fig. 8. Family isotherms of the Langmuir type where $A = 1$ and B changes from 0 to 9.9 (cf. Table 5 for values of B).

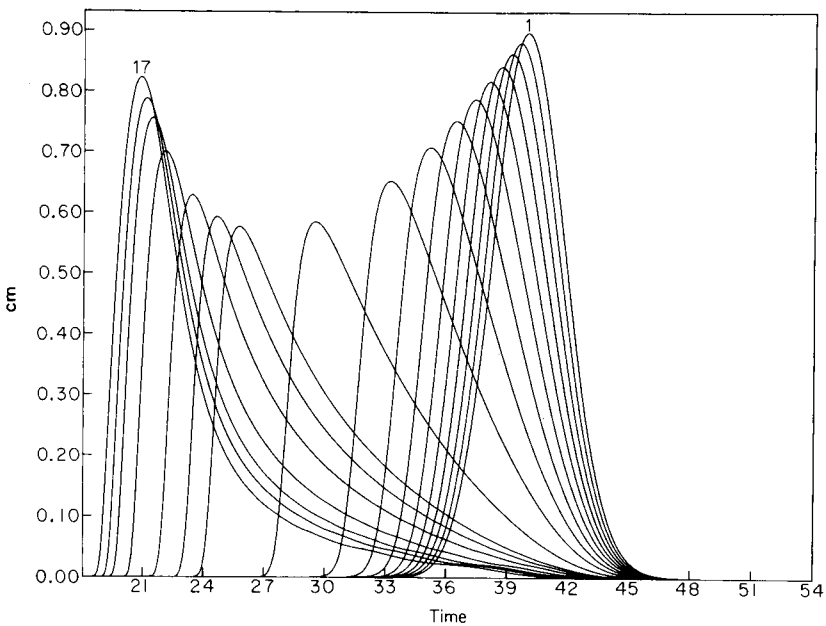


Fig. 9. Family of elution profiles, calculated by the simulation program SAM, showing the influence of the isotherm curvature on the elution profiles; $D = 0.5$ cm^2 s^{-1}; $\langle v \rangle = 5$ cm s^{-1}; $L = 100$ cm; isotherm of the Langmuir type with $A = 1$ and B values as given in Table 5; injection-pulse shaped with $m_0 = 10$; $\Delta z = 1$ cm; $\Delta t = 0.1$ s; with moving coordinates.

TABLE 5

Statistical moments referring to the experimental results presented in Fig. 9

Peak number	B	m_0	μ_1	$\tilde{\mu}_2$	Skewness	Kurtosis
Theor.	0.0000	10.0000	40.080	3.2128	0.1341	3.0300
1	0.0000	9.9990	40.080	3.2127	153.803	3.0795
2	0.0111	9.9988	39.817	3.3133	181.850	3.0650
3	0.0250	9.9988	39.523	3.4457	158.972	3.0577
4	0.0429	9.9987	39.181	3.6255	5.5507	3.0534
5	0.0667	9.9984	38.773	3.8777	180.520	3.0515
6	0.1000	9.9982	38.269	4.2456	180.350	3.0561
7	0.1500	9.9980	37.615	4.8081	167.641	3.0648
8	0.2333	9.9975	36.710	5.7327	150.354	3.0728
9	0.4000	9.9968	35.310	7.4458	124.415	3.0844
10	0.9000	9.9949	32.575	11.4351	90.783	3.1554
11	1.9000	9.9920	29.499	15.9313	75.003	3.4116
12	2.4000	9.9909	28.485	17.0987	15.619	3.5693
13	3.2333	9.9900	27.209	18.0870	18.233	3.8598
14	4.9000	9.9878	25.550	18.1093	22.576	4.4892
15	6.5666	9.9868	24.523	17.1350	27.887	5.1250
16	7.9000	9.9861	23.946	16.1771	31.956	5.6078
17	9.9000	9.9848	23.321	14.7610	14.565	6.2647

CONCLUSIONS AND DISCUSSION

The facilities of the simulation package SAM can be subdivided into two parts. The first part is a stable algorithm which, however, leads to numerical dispersion effects. These effects can be largely reduced by the introduction of moving coordinates. In the case of non-linear isotherms the numerical dispersion increases whenever the maximal curvature of the isotherm increases; for slightly curved isotherms, used mostly in practice, the numerical dispersion is negligible. In cases where a dominant physical dispersion is taken into account the numerical dispersion effect can also be neglected. For sigmoidal isotherms in non-linear mass transport models, the simulation package SAM is applicable in principle, but further mathematical research seems to be necessary in order to check the reliability of the results.

The second part of the simulation package SAM involves the computer program. Besides simulation of pure theoretical problems, the package may also be applied in analytical practice, since experimental isotherms can be used in simulation activities; also experimental elution profiles can be used for recalculating isotherms from peak shapes. Furthermore, almost every subroutine of the second level may be used in other quite different computer applications in analytical chemistry.

The authors thank Drs. R. P. J. Duursma, Dr. T. T. Lub and Dr. Ing. H. Steigstra for their contributions to the general software library used for building the software package SAM.

REFERENCES

1 J. C. Smit, H. C. Smit and E. M. de Jager, Anal. Chim. Acta, 122 (1980) 1 (Part 1).
2 J. C. Smit, Instruction manual simulation package applied to non-linear, non-ideal chromatographic models, in preparation.
3 Ch. Reinsch, Numer. Math., 10 (1967) 177.
4 F. Scheid, Schaum's Outline of Theory and Problems of Numerical Analysis, McGraw-Hill, New York, 1971.
5 E. Kučera, J. Chromatogr., 19 (1965) 237.
6 P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969, pp. 39, 48.
7 H. C. Smit, Chromatographia, 3 (1970) 515.
8 W. F. Ames, Non-linear Partial Differential Equations in Engineering, Academic Press, New York, 1965.
9 J. F. K. Huber and R. G. Gerritse, J. Chromatogr., 58 (1971) 137.
10 A. W. J. de Jong, J. C. Kraak and H. Poppe, J. Chromatogr., to be published.

Analytica Chimica Acta, 122 (1980) 171--178
Computer Techniques and Optimization

PRINCIPAL COMPONENT AND DECOMPOSITION ANALYSIS OF MULTICOMPONENT MIXTURES OF CARCINOGENIC FLUOROPHORES

HARVEY S. GOLD[a], GREGORY T. RASMUSSEN[b], JANET A. MERCER-SMITH[c], DAVID G. WHITTEN and RICHARD P. BUCK*

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

(Received 21st August 1979)

SUMMARY

A series of binary and tertiary mixtures of polycyclic aromatic hydrocarbons is analyzed by using principal component and decomposition analysis of molecular fluorescence spectra. The results demonstrate the ability to determine the number and identity of species that are present. Failures caused by high correlation among spectra are considered.

In recent years, various computer techniques have been applied to chemical systems and problems. Among these have been principal component and decomposition analysis of spectra. Principal component analysis has been described in detail in several books on factor analysis [1—3]. The method has superseded rank analysis as a technique for estimating the number of components in mixtures [4—9] and the literature contains many references to the use of principal component analysis with different kinds of spectroscopic data [10—17]. Likewise, decomposition analysis has recently been presented as a method for analyzing ultraviolet—visible and fluorescence spectra [18, 19]. Both techniques have inherent limitations that can be overcome by utilizing the procedures in concert with each other. In practice, principal component analysis is ideally suited to determining the number of species in a series of related mixtures, but is less capable of identifying these. In contrast, decomposition analysis has been shown to be a powerful method for identifying species in a multicomponent mixture if a reasonable guess as to the number of underlying spectral bands can be made. The estimation of this number can be improved by utilizing the results of principal component analysis. Moreover, the number of species that are present can be further utilized as a constraint in a computer-based identification system for multi-

---

[a]Present address: Department of Chemistry, University of Delaware, Newark, DE 19711.
[b]Present address: Tennessee Eastman Company, Kingsport, TE 37662.
[c]Present address: Wright Chemical Laboratories, Rutgers University, New Brunswick, NJ 08903.

component mixtures that uses decomposition analysis to generate characterizing parameters.

The need to characterize environmental samples increasingly imposes requirements which are not easily met by conventional analytical methods. Illustrative of this problem is the analysis of environmentally significant polycyclic aromatic hydrocarbons (PAH), a class of compounds widely implicated as carcinogenic agents. Separation of a large number of these PAH can be achieved by gas and/or high-performance liquid chromatography (h.p.l.c.), but identification subsequent to separation has proved difficult. In many cases, tedious analyses such as those achieved by h.p.l.c.—mass spectrometry are of little value because of the identical fragmentation patterns of many PAH. The necessity then exists for a technique which can distinguish among PAH and provide data complementary to mass spectrometric analysis. Molecular fluorescence is such a method; its feasibility has been demonstrated by the early work of Freed and Faulkner [20] and of Pellizari and Sparacino [21]. Recent work by Miller and Faulkner [22] demonstrated a method for identifying species from their fluorescence spectra. An alternative scheme developed in this laboratory has the further capability of being able to identify and quantify mixtures of fluorescing species [18, 19]. This is significant in that it allows the requirements for separation of multicomponent mixtures to be relaxed.

In this study the application of principal component analysis to the fluorescence spectra of mixtures is presented. While the discussion is restricted to molecular fluorescence, the results are equally applicable to electronic absorption spectroscopy. Christian et al. [23] have applied methods similar to principal component analysis to fluorescence spectra in the form of an "emission excitation matrix". The present study uses conventional fluorescence emission spectra of related mixtures, in which the relative concentrations of component species vary. The emission spectra of mixtures are assumed to behave as linear combinations of the pure component spectra. A minimum of n mixtures must be available in order to recognize n species as present.

As the number of mixture spectra increases while the number of species is held constant, the system becomes over-specified and the confidence associated with the analysis increases.

If the emission spectrum of any pure component is a linear combination of the other pure components, the number of components present may be underestimated. This is also true if the relative concentrations of pure components in any mixture can be expressed as a linear combination of the concentrations in the other mixtures. In practice, these constraints mean that the method will have difficulty distinguishing species with very similar spectra, and that the number of mixtures used should reflect a reasonable over-estimate of the number of pure components suspected to be present. Such related spectra are quite typical of environmental situations where time-dependent studies are performed to detect changes in individual species

concentrations. Other relevant applications include process control and industrial product control.

In addition to principal component analysis, selected spectra were also analyzed by decomposition techniques. It is significant that species whose spectra were highly correlated proved difficult to analyze satisfactorily by either technique. This is discussed in detail below. The decomposition results are complementary to the previous studies cited above. While accurate determination of the true number of peaks present in a spectrum remains a problem, prior principal component analysis serves to make this far less severe.

EXPERIMENTAL

Excitation and emission fluorescence spectra were obtained by using a Hitachi Perkin-Elmer MPF-2A spectrometer in the direct mode and a Hamamatsu R928 extended red-response multi-alkali (ERMA) photomultiplier tube. Variable emission and excitation slit widths were used. The spectra were digitized at 5-nm intervals.

Spectroquality cyclohexane (MC/B) was used as the solvent without further purification. Heptaphene (Aldrich/Alfred Bader), dibenzochrysene (Aldrich/Alfred Bader), dibenz[*a*,*h*]anthracene (Eastman reagent grade), pentacene (Aldrich reagent grade), and perylene (Aldrich Gold Label) were quantitatively dissolved with ultrasonic shaking without further purification. Initial stock solution concentrations were 123, 165, 236, 27.2, and 202 $\mu g\ ml^{-1}$, respectively.

DATA ANALYSIS AND RESULTS

The intensity values of the digitized spectra were computer-adjusted to eliminate effects of variable slit widths. Decomposition analysis was done by using the program SPECSOLV as discussed in previous papers [18, 19, 24]. SPECSOLV is a generalized algorithm designed to perform decomposition (sometimes termed "spectral-stripping" and erroneously called "deconvolution") of spectra composed of overlapped symmetric or asymmetric Gaussian peaks. The spectrum is baseline-adjusted and then transformed to a discrete function of data points at constant wavenumber increments by interpolating values at intermediate points on a cubic segment passing through the four nearest input points. Spectral analysis proceeds iteratively, resolving the spectrum into component peaks. After a preliminary search, initial parameters are obtained. Theoretical peaks calculated from these parameters are subtracted from the spectrum. An iteration consists of adding, searching for, and subtracting each calculated peak from the residual spectrum. After each iteration, the sum of the squared residuals is compared with a tolerance factor. If, after a reasonable number of iterations, the sum has not approached a minimum, processing of the data set is terminated.

Calculated parameters consist of peak locations, relative intensities, and half-width parameters.

The principal component analyses were done by standard techniques; for each set of mixed spectra, the data were treated as an l by m data matrix, D, where l is the number of wavelengths at which spectra were digitized and m is the number of mixed spectra. A matrix element, d_{ij}, corresponds to the fluorescence emission intensity observed at the i^{th} wavelength on the j^{th} mixture. A second moment matrix (covariance about the origin), C, was computed and the eigenvectors, $\vec{e}$, and eigenvalues λ of this matrix were calculated: $C = D^t D$ and $C\vec{e}_k = \lambda_k \vec{e}_k$.

Each eigenvector represents a principal component of the data and the associated eigenvalue reflects the variance spanned by that vector. The eigenvalues, arranged in descending order, are used to estimate the number of true components. For ideal data, the eigenvalues corresponding to true components are positive and all others are zero. With real data, all eigenvalues may be positive so that the problem is to determine which eigenvalues correspond to true components, and which eigenvalues correspond to components arising from error or noise in the system. Various criteria to determine the number of true components have been reported [25, 26]. Two such were compared in this study. One criterion is to select the n largest eigenvalues necessary to include 99% of the total variance. The other criterion is based on an empirical function described by Malinowski [26], which considers the magnitudes of the eigenvalues and the dimensions of the original data matrix. Both criteria were used with all sets of mixed spectra.

Several binary and tertiary mixtures of A, B, and C were prepared and fluorescence emission spectra were obtained for each. Concentrations for the twenty-four mixtures are listed in Table 1. The results of the principal component analyses for the five sets of mixtures are summarized in Table 2. The mixtures of A and B, and those of B and C, were correctly identified as containing two components by both criteria. The mixtures of A and B were also used to test the case where the concentration of one component is constant while the concentration of another varies. For the set of four AB mixtures having a constant concentration of A, both criteria indicate a one-component "mixture". However, other sets of four AB mixtures are estimated correctly to be two-component mixtures. Thus, for a set of mixtures in which the concentration variation on one or more components is insufficient, the principal component analysis may underestimate the number of species present. A related problem is encountered in the analysis of the two sets of AC mixtures. Here the spectra of the two pure compounds are so similar that the principal component analysis cannot recognize two components. Both criteria indicate only one component for the first set of mixtures whereas, for a set of less concentrated mixtures, the two criteria disagree about the number of components present. If the spectra of the independent pure components are too similar, the principal component analysis may again underestimate the number of components. This explains the

TABLE 1

Concentrations of samples in $\mu g\ ml^{-1}$ with cyclohexane as solvent

Sample		Constituents[a]		
Number	Type	A	B	C
1	AB	0.0011	61.50	
2		0.0165	15.40	
3		0.0165	20.30	
4		0.0164	41.20	
5		0.0164	61.50	
6	BC		15.40	0.0236
7			20.30	0.0236
8			41.20	0.0236
9			61.50	0.0158
10			61.50	0.0236
11	AC 1	0.033		15.70
12		5.500		7.86
13		8.250		12.80
14		11.000		7.86
15	AC 2	0.00330		0.00472
16		0.00330		0.00118
17		0.00660		0.00590
18		0.00825		0.00472
19		0.00990		0.00354
20	ABC	0.00825	30.75	0.0158
21		0.00825	30.75	0.0236
22		0.00825	41.20	0.0118
23		0.00825	61.50	0.0118
24		0.01650	30.75	0.0158

[a]A, B and C represent dibenzochrysene, heptaphene and dibenz[*a*,*h*] anthracene, respectively.

TABLE 2

The effect of different sample subsets upon the predicted number of components

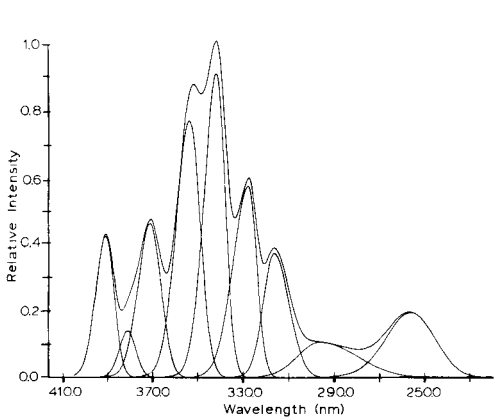
Sample type[a]	Sample number	Estimated number of components from	
		Malinowski function	99% variance
AB	1—5	2	2
AB	1—4	2	2
AB; A constant	2—5	1	1
BC	6—10	2	2
AC 1	11—14	1	1
AC 2	15—19	2	1
ABC	20—24	2	2

[a]A, B and C represent dibenzochrysene, heptaphene, and dibenz[*a*,*h*]anthracene respectively.

results observed for the ABC mixtures in that two components are predicted since A and C were not recognized as independent species by the analysis.

Spectral decomposition results are consistent with the principal component analysis. Identification of peaks by reference to the characterized standard spectra is generally satisfactory. All of the samples studied possess a large number of component peaks. Figure 1, illustrating the successful decomposition of the dibenz[*a*,*h*]anthracene excitation spectrum, is representative. (During the study, this species was designated as C, while dibenzochrysene and heptaphene were A and B, respectively.) Unambiguous identification of mixtures was possible with the exception of the $A_n C_m$ combinations. In these instances, convergence was obtained, but frequently this was upon spurious peak parameters. The inability to resolve $A_n C_m$ mixtures is due to the high degree of correlation in the spectra of these materials. This is illustrated graphically in Fig. 2 where the spectra of all three species are overlaid. Clearly, the dibenz[*a*,*h*]anthracene and dibenzochrysene spectra are highly correlated, while heptaphene is quite visibly different.

The degree of correlation is related to the ability of principal component and decomposition methods to distinguish species properly. Yet the effect differs in several ways. In the former, high correlation causes the species to appear to be a single species. In the latter, decomposition of single-species solutions results in unequivocal compound identification, yet the method fails properly to identify closely correlated spectra of species present in mixtures. At this point, the relative effects of correlation of the spectral envelope versus correlation of component peaks are uncertain.



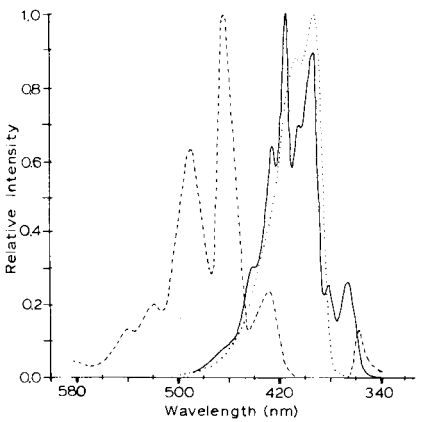


Fig. 1. Result of the computer decomposition of the excitation spectrum of dibenz[*a*,*h*]-anthracene. Nine component peaks are present. The relative intensity is assigned to unity for the most intense component peak.

Fig. 2. Superimposed emission spectra of dibenz[*a*,*h*]anthracene (——), dibenzochrysene (·····) and heptaphene (------), illustrating the high degree of correlation of the first two species.

Thus correlation of spectra leads to ambiguity in both methods, but with strikingly important differences. In the case of two or more species with uncorrelated spectra as a mixture, principal component analysis will find the correct number of species. These data can be used in turn to speed identification by decomposition analysis. In the case of two or more species with highly correlated spectra as a mixture, principal component analysis finds too few species to be present. This number, or even the correct value, is utilized in a decomposition analysis. The result is likely to be convergence upon spurious peaks that probably belong to no single species or reasonable combination of species. The method still retains the ability to identify the species after separation, with the need for separation indicated by simultaneous consideration of the results of both analyses.

Figure 3 shows the successful decomposition of the emission spectrum of solution 4, a mixture of heptaphene and dibenzochrysene in cyclohexane. Comparison of peak parameters with those obtained for the individual reference species in cyclohexane indicates good agreement and unambiguous identification. Similar results are obtained with other mixtures.

CONCLUSION

The two techniques of principal component and decomposition analysis can be used together to advantage. Problems arising from close correlation of spectra are apparent in both, and can be recognized by proper data analysis. It is likely that proper selection of criteria used to judge the number of components can reduce the problem; in some instances, use of the Malin-

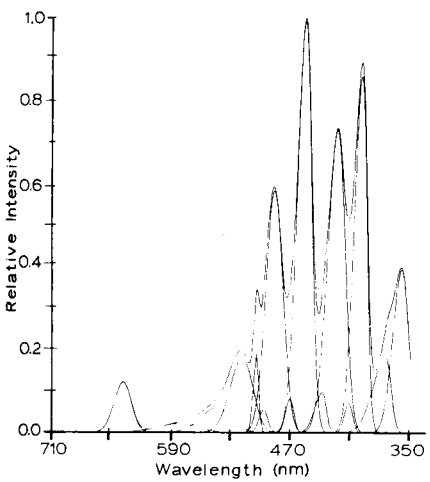


Fig. 3. The successful decomposition of a binary mixture of heptaphene and dibenzochrysene in cyclohexane. Peak parameters permit identification of the species by reference to a library file of characterized spectra.

owski criterion provides better answers, yet judgement of the effects of correlation on the course of both numerical techniques is as yet empirical and not quantifiable or predictable.

Partial support for this work by the National Science Foundation under Grants MPS-75-00970, CHE 77-14547 and CHE 77-2049 is gratefully acknowledged.

REFERENCES

1 R. J. Rummel, Applied Factor Analysis, Northwestern University Press, Evanston, IL, 1970.
2 P. Horst, Factor Analysis of Data Matrices, Holt, Rinehart and Winston, New York, 1965.
3 R. J. Harris, A Primer of Multivariate Statistics, Academic Press, New York, 1975; H. H. Harmon, Modern Factor Analysis, University of Chicago Press, Chicago, 1970.
4 R. M. Wallace, J. Phys. Chem., 64 (1960) 899.
5 G. Weber, Nature (London), 190 (1961) 27.
6 S. Ainsworth, J. Phys. Chem., 65 (1961) 1968.
7 R. M. Wallace and S. M. Katz, J. Phys. Chem., 68 (1964) 3890.
8 D. Katakis, Anal. Chem., 37 (1965) 876.
9 L. F. Monteiro and R. I. Reed, Int. J. Mass Spectrom. Ion Phys., 2 (1969) 265.
10 J. J. Kankare, Anal. Chem., 42 (1970) 1322.
11 Z. Z. Hugus and A. A. El-Awady, J. Phys. Chem., 75 (1971) 2954.
12 W. H. Lawton and E. A. Sylvestre, Technometrics, 13 (1971) 617.
13 N. Ohta, Anal. Chem., 45 (1973) 553.
14 J. T. Bulmer and H. F. Shurvell, J. Phys. Chem., 77 (1973) 256.
15 J. E. Davis, A. Shepard, N. Stanford and L. B. Rogers, Anal. Chem., 46 (1974) 821.
16 J. McK. Halket and R. I. Reed, Org. Mass Spectrom., 10 (1975) 808.
17 G. L. Ritter, S. R. Lowry, T. L. Isenhour and C. L. Wilkins, Anal. Chem., 48 (1976) 591.
18 H. S. Gold, C. E. Rechsteiner and R. P. Buck, Anal. Chem. Acta., 95 (1977) 51.
19 H. S. Gold, C. E. Rechsteiner and R. P. Buck, Anal. Chem. Acta., 103 (1978) 167.
20 D. J. Freed and L. R. Faulkner, Anal. Chem., 44 (1972) 1194.
21 E. D. Pellizari and C. M. Sparacino, Anal. Chem., 45 (1973) 378.
22 T. C. Miller and L. R. Faulkner, Anal. Chem., 48 (1976) 2983.
23 G. D. Christian, I. M. Warner, E. R. Davidson and J. B. Callis, Anal. Chem., 49 (1977) 564.
24 H. S. Gold, C. E. Rechsteiner and R. P. Buck, Anal. Chem., 48 (1978) 1540.
25 D. L. Duewer, B. R. Kowalski and J. L. Fasching, Anal. Chem., 48 (1976) 2002.
26 E. R. Malinowski, Anal. Chem., 49 (1977) 612.

Analytica Chimica Acta, 122 (1980) 179—191
Computer Techniques and Optimization

A VERSATILE COMPUTERIZED SYSTEM FOR THE DEVELOPMENT AND COMPARISON OF ELECTROANALYTICAL PROCEDURES

HANS JØRGEN SKOV and LARS KRYGER*

Department of Chemistry, Aarhus University, Langelandsgade 140, DK-8000 Aarhus C (Denmark)

(Received 13th August 1979)

SUMMARY

The computerized system reported is suitable for potentiometric and voltammetric techniques. All instrumental settings are fully automated and new analytical schemes can be implemented simply by software modifications. A dedicated 4K, 16-bit microcomputer allows rapid real-time data acquisition and processing. By application of a data acquisition scheme simulating a multichannel analyser, stripping potentiograms and chronopotentiograms can be conveniently represented in digital form as time vs. potential relationships. Procedures developed on the system can be transferred to microprocessor-controlled equipment as the software for complete analytical procedures, including device handlers, data storage buffers and graphic output, occupies 3—4K words. Flexibility in program editing and assembling is obtained with an optional data link to a medium-sized, time-shared computer with extensive software packages. The link described is of simple construction and can be readily established between computers with standard teletype terminals. The performance of the system is illustrated by comparing the selectivities of two established voltammetric stripping techniques with the selectivity of a new potentiometric technique developed on the system.

During recent years electroanalytical techniques have been developed very extensively. Several of these techniques (e.g. stripping analysis, differential pulse polarography and amperometric detection in liquid chromatography) are characterized by high sensitivities and are powerful tools in trace analysis.

The advantage of employing computer automation in conjunction with electrochemical and in particular voltammetric trace analysis has been demonstrated by several workers [1—7]. Not only may computerization of the electrochemical experiment lead to improved precision but also, in favourable cases, real-time on-line computer interaction with the experiment may improve the selectivity [7] and accuracy [5] of a technique.

In this laboratory computerized electrochemical techniques of trace analysis and their evaluation are under study [5, 6]. Similar studies are being carried out at a number of laboratories [4, 8, 9] but so far these projects have been concerned with comparisons of voltammetric and modified voltammetric techniques, the main interest being concentrated on various types of a.c., square-wave and pulsed techniques.

It has recently been demonstrated that potentiometric techniques are suitable for trace element determinations at, and even below, the $\mu g\ l^{-1}$ level. Potentiometric stripping analysis (p.s.a.) as reported by Jagner [10] and Jagner and Graneli [11] is a promising technique with analytical characteristics comparable with those of the voltammetric techniques. The p.s.a. technique has found several applications but only one modified technique, multiple scanning potentiometric stripping analysis (m.s.p.s.a.) [6], has so far been reported. Furthermore, the entire field of elucidating the potentiometric stripping process, in particular by digital simulation, has not been explored. It seems likely that the potentiometric techniques will undergo a development similar to the voltammetric techniques and there is little doubt that for certain problems (e.g. involving media of low ionic strength) potentiometric techniques will turn out to be superior to the voltammetric ones.

This paper describes a fully computerized system for the development and comparison of electroanalytical procedures. As potentiometric techniques of trace element analysis will soon become commonplace, the system developed, apart from the functions commonly available in computerized voltammetric equipment, provides for a potentiometric mode so that direct and rapid comparisons between voltammetric and potentiometric experiments on the same sample are feasible.

The 4K microcomputer used is sufficiently large to act as a stand-alone computer and simultaneously hold programs for both voltammetric and potentiometric experiments. Program development, i.e. editing and assembling, takes place in a medium-sized, time-shared, departmental computer with extensive software and backing storage facilities. Once assembled, the programs are transferred to the microcomputer. Moreover, experimental data may be moved from the micro to the medium-sized computer for comparison with theoretical data. The computer intercommunication is via a common teletype line, and is almost entirely based on software which allows the microcomputer to simulate a teletype. This scheme makes the system well suited to studies involving both experimental and theoretical aspects. Because of the simplicity of the computer-link, the computerized electrochemical system is completely detachable from the departmental computer and may be moved from laboratory to laboratory. Only during program development, or when experimental data are to be delivered to the departmental computer, is the system connected to the nearest teletype line. Thus, although the combined system provides the full computational power of the departmental computer, experiments are not affected by the drawbacks of time-sharing and maintenance periods typical for departmental installations.

SYSTEM DESIGN CONSIDERATIONS

Data acquisition mode and fast multichannel potentiometry

Rapid real-time response is of great importance to the sensitivity and precision of computerized voltammetric techniques. This is also true for

potentiometric techniques of trace analysis [6]. Although potentiometric stripping analysis for elements at or below the $\mu g\ l^{-1}$ range can be carried out with very simple equipment [10] if the time allowed for electrolytic preconcentration is sufficiently long (30 min or more), the introduction of a fast-responding computer allows the monitoring of the transient stripping signals occurring even after short plating periods. Furthermore, a fast digital computer provides a particularly convenient mode of sampling time data like stripping potentiograms (or chronopotentiograms) where the analytical signal as a function of potential is the total time during which the working electrode experiences that potential.

The data acquisition situation is similar to that encountered when the energy distribution of radiation quanta is to be recorded. However, where radiation quanta arrive in a random sequence because of stochastic processes in the system under study, the cell-potential readings may be obtained equally spaced in time by the use of an A/D converter triggered by a clock pulse-generator.

A device similar to a multichannel analyser is therefore suitable for recording digitized stripping potentiograms and chronopotentiograms: each channel represents a small potential interval. Whenever a potential reading is available, the count in the corresponding channel is incremented. Data acquisition terminates after a preset time, or if a reading falls outside a preset potential range. The assignment of channel number to a given potential reading can be done by analog circuits. However, a small digital computer may act as a multichannel analyser and, by processing data in real time, may assign the proper core address to any potential reading and subsequently increment that core location. By using this scheme, the potentiometric data are represented in the natural way, i.e. potential is the independent variable and time the dependent variable (the analytical signal). This is not the case if a simple series of potential readings are taken at regular time intervals. The point-by-point addition of "multichannel potentiograms" is therefore straightforward and "multichannel potentiometry" thus suitable for signal-averaging techniques which may improve sensitivity and precision considerably [6].

In the computerized system described here, the multichannel facility has been implemented in the software. As the precision (counting statistics) is closely related to the real-time accuracy of the system, care has been taken to ensure that this quantity is maximized by employing a rapid clock generator and assembler programming. A system of maximum real-time accuracy, moreover, permits software simulation of analog electronic operations. Thus, for example, differential pulse polarography can, with suitable software, be run without any analog signal integration and subtraction. When a technique is eventually optimized, the simulation algorithms may be partly replaced by equivalent electronic circuits to produce a compact analytical device.

Computer size

While computerized electroanalytical techniques are very demanding with respect to real-time accuracy, their requirements of data storage are less critical. In voltammetric and potentiometric procedures, resolutions better than 1 mV are difficult to obtain and often 5 mV is sufficient. Moreover, as common electrode materials impose a practical limitation of 1—2 V on the potential range which can be studied, data buffers for voltammograms rarely exceed 1K words. Even for time data such as chronopotentiograms, data buffers of 1K words are usually sufficient. This is true because in electrochemical trace analysis chemical factors rather than limitations imposed by instrumental design govern the resolution.

Complete electroanalytical programs including device handlers and data storage buffers therefore require typically 3—4K words of computer memory and do not normally need any external data storage capacity. This is fortunate, because practical analytical chemical work frequently requires portability of equipment. From a small, dedicated computer, the translation of analytical procedures to other systems, e.g. microprocessor-controlled apparatus for field work, is simple because the instruction sets for most small computers/processors are very similar.

HARDWARE CONFIGURATION

A block diagram of the hardware configuration is shown in Fig. 1.

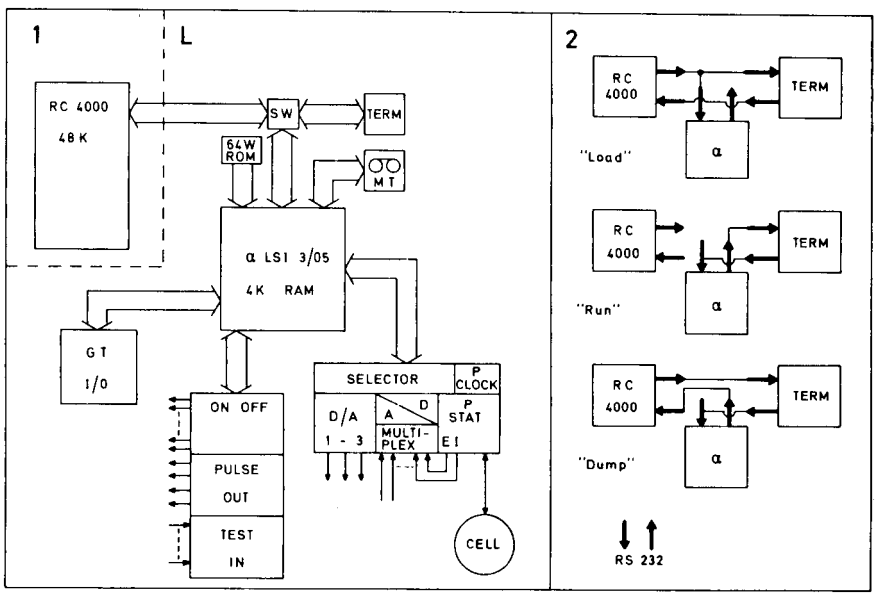


Fig. 1(1) System hardware configuration and 1(2) data link to departmental computer.

Microcomputer

The microcomputer employed is an ALPHA LSI-3/05 naked version (4K words of 16-bit RAM). Typical execution times range from 4 to 10 μs for most instructions. Direct memory access is possible with the so-called automatic I/O instruction set (approximately 20 μs). The computer, which is commercially available as two print-cards, may communicate with a maximum of 16 external devices on an interrupt basis.

A system clock with a fixed frequency of 100 Hz is an integral part of the naked computer version. This clock, which is associated with a special interrupt address, may be used for low frequency data acquisition and control in real time.

Terminal and communication with larger computer

An ADM-3A alphanumeric terminal may communicate with the microcomputer through standard RS232 interfacing. Via the switchboard (SW), the function of which is shown in greater detail in Fig. 1(2), any computer with standard teletype I/O facilities may take part in this conversation. In the work presented here, this computer was an RC4000, 48K 24-bit computer with extensive peripherals including 2 discs (2 and 8M words) and an 8-track magnetic tape station.

The RC4000 computer in turn acts as a terminal for a CYBER 172 and complex data processing and comparisons of experimental data with simulated data can be conveniently run on either of those computers. Three typical situations can be handled by this circuitry in Fig.1(2).

Load situation. This situation allows the microcomputer to copy information sent to the alphanumeric terminal by a larger computer. To move programs and data from the larger computer to the microcomputer, the operator simply initiates copying of the relevant source files from the larger computer to the terminal. A 64-word loader program, which may be loaded into the high core area of the microcomputer from a 64-word ROM, simultaneously interprets the serial ASCII information as hexadecimal code and stores it sequentially from a given core address.

A cheap cassette tape-recorder (MT) is interfaced to the microcomputer and runs in parallel with the alphanumeric terminal. In this way, hard copies of the conversation between the two computers may be obtained, i.e. programs and data may be stored on and read from magnetic tape.

Run situation. Here the system is completely detached from the larger computer. This situation is the one most commonly used when the program development and debugging phase has been completed. After "power-down" program reloading is necessary (mostly from tape) as the microcomputer memory is volatile.

Dump situation. In this mode the microcomputer software simulates the actions of an operator using an alphanumeric terminal for input to the larger computer. Full duplex communication is used, i.e. whenever a character has been input to the RC4000, the microcomputer expects and waits for an

identical character to be echoed from the large computer. The rate of data transfer between the two computers depends solely on the teletype line used (10—100 ASCII characters per second). This slow rate is not crucial, since the microcomputer usually runs independently, and the link is not needed for any real-time operations.

The real advantage of the scheme is that in order to establish a link from the microcomputer to any other computer with teletype I/O, no hardware and only slight software modifications are necessary.

Graphical displays

Raw or processed data may be displayed on a dynamic graphical display [12] modified to include graphical input (cursor positions) or on a Philips PM8041 x—y recorder.

Binary state module

This module controls and monitors binary-state devices such as stirrer motors, syringe burettes, microswitches and valves for gas flow. It comprises 12 computer-controlled TTL on/off relays. The output from each of these relays may be converted into 220 V a.c. on/off functions. Furthermore, four single-pulse outputs (5 V), e.g. for syringe-burette control, are available. The pulse durations (1—1000 ms) may be set individually on front panel knobs. Finally, the state of the binary-state devices may be monitored through six test inputs which are sensitive to short-circuiting. These inputs are mostly used to indicate whether injection loops for flow analysis are completely open or completely closed.

Analog signal controller/monitor

This device can be regarded as four separate sub-devices, i.e., the D/A converter unit, the A/D converter unit, the programmable clock and the electrochemical module. A sub-address field (function code) of any I/O instruction is decoded by a selector and gates the data flow to or from the appropriate sub-device (see Table 1).

D/A converter unit. Analog voltages (±5 V, 12-bit resolution) are available from two D/A converters. These converters are normally used for controlling the Philips PM8041 x—y recorder. A third 10-bit D/A converter supplies currents of 1—1000 μA and is used for chronopotentiometry.

A/D converter unit. The 12-bit A/D converter used has a range of ±2 V and a conversion time of 25 μs. The converter has 16 analog inputs, two of which are normally connected to the potential- and current-amplifier outputs (E and I) of the electrochemical module. The A/D conversions are triggered by the programmable clock. By means of automatic analog multiplexing, cyclic monitoring of the analog inputs is possible. The multiplexer is synchronized with the programmable clock, and any monotonic subset of the input channels starting at channel 0 may be included in the cycle. This feature is particularly useful in voltammetric applications, since automatic

TABLE 1

List of function codes for analog module. (The second column refers to digital inputs to the module and electronic switches as shown on Fig. 2.)

Function Code	Input	Function initialized
0	7	Frequency of programmable clock, 0.25 Hz—1 mHz
1	*g*	Frequency of pulse train to be superimposed on potential ramp. The pulse train is a subset of the programmable clock pulse train
2	*g*	Potential pulse duration in units of programmable clock pulses
3	5	Potential pulse amplitude, —1000→ +1000 mV. Resolution 0.5 mV
4	*i, j, k, l* 6	Current range/current bias selection. Available ranges 1 μA, 50 μA, 200 μA, 1 mA
5	*a, b*	Potentiostat resting potential selection. Upper (anodic) or lower (cathodic) as specified by function codes 6 and 7
6	1	Upper (anodic) resting potential, —2000→ +2000 mV, resolution 1 mV
7	2	Lower (cathodic) resting potential, —2000→ +2000 mV, resolution 1 mV
8	3	Slope of anodic potential ramp. Minimum 1 mV s^{-1}, maximum 1024 V s^{-1}
9	4	Slope of cathodic potential ramp. Minimum 1 mV s^{-1}, maximum 1024 V s^{-1}
10		DAC1 —2500 mV → +2500 mV, 12-bit[a]
11		DAC2 —2500 mV → +2500 mV, 12-bit[a]
12		DAC3 —2500 mV → +2500 mV, 12-bit[a]

[a]Not shown on Fig. 2.

switching between potential and current readings produces a more accurate voltammogram than one obtained by computing the potential as a function of ramp generator calibration results and time.

Programmable clock

For multichannel operation in potentiometry and rapid-scanning voltammetry, the data acquisition controlled by the computer real-time clock is too slow (max. 100 Hz). Therefore, a programmable clock with a base frequency of 1 MHz has been added to the system. Two ranges are available, 1—4095 μs and 1—4095 ms, but of course the maximum data rate is determined by the A/D conversion rate.

Electrochemical module

The electrochemical module is designed with potentiometric, voltammetric, and amperometric experiments in mind, and the selection of the potentiometric or voltammetric mode is under computer control. Control functions such as potential ramp generation, possibly with pulse or square-wave

modulation, are fully automated and synchronized with the programmable clock, which also triggers the A/D converter. Also, the selection of current range and bias of the voltammetric unit is under program control.

Table 1 summarizes the computer-controlled functions of the electrochemical module, a block diagram of which is shown in Fig. 2. The module comprises a potential ramp generator and a potentiostat with potential-pulse generation facilities. The anodic and cathodic resting potentials of the potentiostat are set via D/A converters 1 and 2 (all converters and registers in Fig. 2 having 12 bits). Selection of anodic or cathodic resting potential is done via the electronic switches *a* and *b*. The potential ramp is supplied by the integrating circuits below D/A converter 2 and may be added to the resting potential by closing the electronic switch *c*. The integration is activated via the electronic switch *d*. Two voltage ranges may be supplied to the integrator: in rapid scanning voltammetry, switch *e* is open while switch *f* is closed and vice versa for slow scanning. The minimum scan rate is 2 mV s^{-1} and the maximum is 1000 V s^{-1}, although the potentiostat in its present version is not suitable for potential scans as fast as 1000 V s^{-1}, as provision for IR compensation has not been made. Two registers (3 and 4) hold the digital representation of the anodic and cathodic scan rates, respectively.

By means of the ramp generator potential steps, linear potential sweeps and cyclic scans (asymmetric if desired) can be created. The resultant potential is fed to the potentiostat and may be superimposed by a pulse train (or

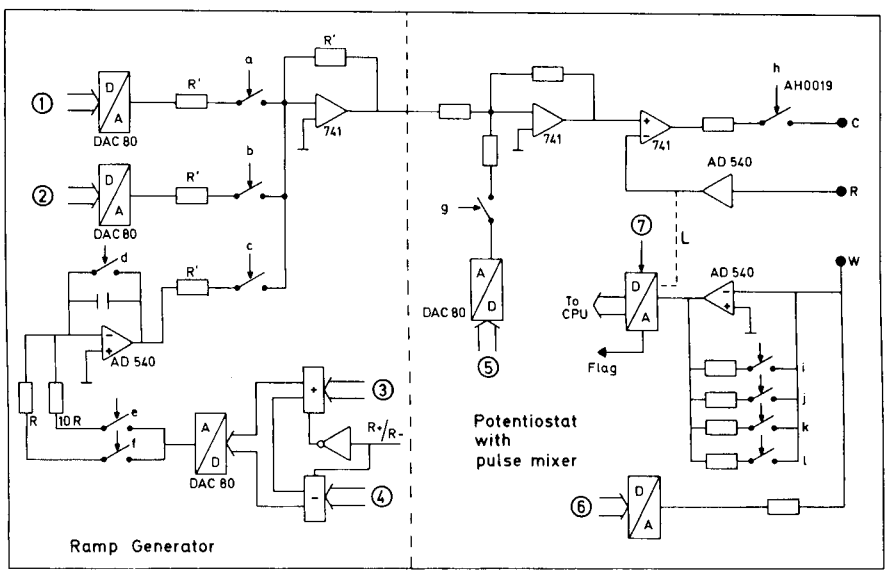


Fig. 2. Electrochemical module comprising ramp generator and potentiostat with pulse mixer.

square wave). The amplitude of this modifying potential is set by D/A converter 5. The pulse duration and frequency are synchronized with the programmable clock and are activated via switch *g*. For a given input voltage the potentiostat maintains a constant potential between the reference and working electrode terminals *R* and *W*. In the voltammetric/amperometric mode, switch *h* is closed and current may pass through the counter electrode at *C*. The current through the working electrode may be monitored by using one of four different current ranges (switches *i*, *j*, *k* and *l*). Via D/A converter 6, a biasing current may be employed. The resultant current is converted to a voltage and fed to the A/D converter.

When the module is used for potentiometric measurements, switch *h*, a reed-relay with 0.1-ms switch time, is opened and the potential of the reference electrode is fed to the A/D converter via the dotted line *L*. Throughout the electrochemical module, analog and digital signals are coupled optically in order to minimize noise.

SOFTWARE

In order to obtain a computerized system where new analytical schemes are easily tested, software modifications must be simple. High-level language programming [13—15] offers a number of advantages in this respect, and program packages (BASIC, FORTRAN, ALGOL, etc.) with special assembler-coded routines for rapid I/O of blocks of data to/from experimental devices are commercially available for most computers. Although such specialized I/O routines can be directly accessed by high-level language calls, they are usually not suitable for rapid intelligent interaction with the experiment in real-time, e.g. the multichannel analyser approach. The reason for this is that the access time of high-level language statements is relatively long and difficult to predict. Thus, if the computer should respond intelligently during input of a block of data (e.g. in multichannel potentiometry, to re-establish potentiostatic control if the anodic potential limit is reached) the pure high-level language solution would require an exit from the I/O routine and access of statements of the "IF - - - THEN" type, which in turn would decrease the real-time accuracy to typically several milliseconds [3, 14].

Authors who use high-level programming for rapid real-time interaction therefore always report extensions of the language with assembler-coded subroutines to suit their specific purpose. In fact, the experimentalist who wants to run a closed-loop experiment, where the course of data collection is updated as a function of rapidly appearing observations, must necessarily familiarize himself with assembler programming and machine language.

In recognition of this fact, assembler programming only is used here [16]. This strategy also greatly reduces the necessary computer memory. Furthermore, if an analytical procedure is described as a sequence of machine instructions, the implementation of an electronic device such as a microprocessor with equivalent performance is simpler.

TABLE 2

Core lay-out

Page number	Hexadecimal address	Contents
0	0000—002F	External interrupt locations and pointers for interrupt service routines
	0030—003F	Software flags for pending interrupts
	0040—0060	Subroutine address pointers
	0061—007F	Constants
1—22	0080—0093	Internal interrupt locations
	0094—00B8	Current experimental parameters
	00B9—00BC	Data buffer addresses
	00BD—00CB	Internal interrupt handling routines
	00CC—00F8	Auto I/O end of block handling
	0112—0233	Alphanumeric terminal I/O handler. Preparation and interpretation of character string
	0234—0309	Routines for conversions between binary/decimal/ hexadecimal representations. Character I/O to/ from alphanumeric terminal
	030A—0384	Integer arithmetic — add, subtract, integrate, etc.
	0385—0395	Data buffer address initialization
	0396—0419	Graphic terminal handler
	041A—060C	Handlers for experimental devices
	060D—0726	User block: programs comprising mainly sub-routine calls
23—31	0C00—0FFF	Four data buffers

All program editing and assembling has been carried out on the RC4000 computer.

Core lay-out

The core lay-out is shown in Table 2. The zero page, which is the directly addressable part of the computer memory (128 words) is reserved for interrupt locations, address pointers for indirect addressing, constants and a number of flag locations (one per external device), the contents of which indicate whether the corresponding device is idle or an interrupt is pending. These flags are set by the running program prior to initiating input/output to or from external devices. Whenever a device has accomplished its action, the running program is automatically interrupted by the execution of an interrupt service routine, which clears the flag. This scheme enables the running program to detect whether data acquisition and control takes place in real-time. If, for example, during input of a series of digitized data, the program senses a reset flag prior to a datum-input operation, the program loop is too slow to handle the A/D conversion frequency used and a real-time error has occurred. If data acquisition takes place in real-time, the program has to wait for input since the interrupt is still pending and the flag still set when the program is ready for input.

Most of the program consists of a fixed set of system subroutines. New operational modes are implemented by adding a block of typically 100 instructions. These instructions are mainly subroutine calls and algorithms for real-time interaction with the experiment. At present such program blocks have been written for stripping potentiometry, stripping voltammetry, differential pulse stripping voltammetry, square-wave voltammetry, staircase voltammetry and cyclic linear-sweep voltammetry.

Apart from four data buffers in high core, a data area is reserved for storage of experimental parameters such as programmable clock frequency, potential scan-rate, current measurement range, scale factors for graphic displays, etc. These parameters may be thought of as "knobs" on the front panel and are set prior to the experiment by typing them on the keyboard.

Having initiated all D/A converters and the programmable clock, the program may select and start the relevant functions by outputting a special select-instruction to the analog module. The data carried by this instruction are a sum of codes as shown in Table 3. Thus, for example, a single potential sweep with alternate current and potential readings through channels 0 and 1 would require that 1 + 32 + 64 + 256 = 353 be output by the select instruction. The exerimenter communicates with the system by entering mnemonic command codes possibly followed by data from the keyboard.

RESULTS AND DISCUSSION

Recently, a new technique, multiple scanning potentiometric stripping analysis (m.s.p.s.a.) was developed by using the system reported here [6]. As in potentiometric stripping analysis (p.s.a.) [10, 11] the analytical signal of a component determined with m.s.p.s.a. is proportional to the time used to redissolve the preconcentrated component from the working electrode during the stripping step, but in m.s.p.s.a. the analytes, once preconcentrated, are forced to undergo several chemical oxidation/potentiostatic reduction cycles. During the oxidation parts of these cycles, the computer acquires and adds the analytical signals, i.e. the number of time units (clock pulses) spent within any potential interval of the pre-selected potential window to be studied. In this manner a considerable signal enhancement is obtained. The data acquisition scheme in m.s.p.s.a. has previously been discussed [6].

To illustrate further the performance of the system, a comparison of selectivities was carried out between m.s.p.s.a., anodic stripping voltammetry (a.s.v.) and differential pulse anodic stripping voltammetry (d.p.a.s.v.). In a.s.v. overlap problems between adjacent peaks are likely to occur at high potential scan rates [17], particularly if the sample solution contains a large concentration of the substance which is redissolved at the more cathodic potential. Because a rapidly increasing potential is imposed on the working electrode, such a substance may not be completely stripped when the stripping of the more anodic component begins. The problem may be solved by going to slow potential scans, but this in turn leads to decreased sensitivity.

TABLE 3

List of select codes for analog module

Code	Action
1	Single potential sweep, voltammetric monitoring
3	Multiple cyclic potential sweeps, voltammetric monitoring
7	Single cyclic potential sweep, voltammetric monitoring
8	Potentiometric monitoring, potentiostat off
16	Potential ramp superimposed by pulse train
32	Cyclic monitoring of n analog inputs. $0 \leqslant n \leqslant 16$
64	A/D conversion series triggered by programmable clock
128	Single A/D conversion
$n \times 256$	Analog input selection; channels 0 through n are scanned in a cyclic manner

In p.s.a. (and m.s.p.s.a.) the analytes are chemically redissolved one after another, and the overlap problem should be less severe.

To compare the selectivities of a.s.v., d.p.a.s.v. and m.s.p.s.a., solutions containing 200 ppb cadmium(II) chloride, 10 ppb lead(II) nitrate, and 100 ppm mercury(II) acetate in 0.5 M sodium chloride were electrolysed for 60 s at −1300 mV vs. SCE at a mercury-film glassy carbon working electrode. The results of the subsequent stripping experiments are shown in Fig. 3.

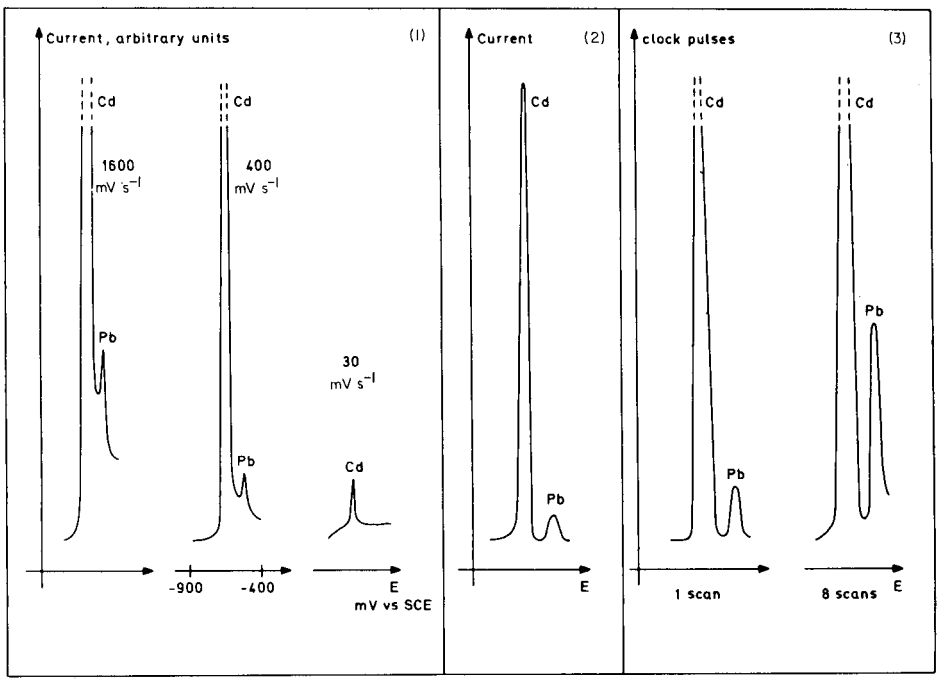


Fig. 3. Comparison of selectivities in (1) a.s.v., (2) d.p.a.s.v., and (3) m.s.p.s.a. of a solution of 200 ppb cadmium(II) chloride, 10 ppb lead(II) nitrate and 100 ppm mercury(II) acetate in 0.5 M sodium chloride.

The linear sweep voltammograms clearly show the interference of cadmium with lead. The interference may be reduced by going to lower potential scan rates, but this results in decreased peak heights and at 30 mV s^{-1} the lead peak is no longer visible. The d.p.a.s.v. experiment was run at a potential scan rate of 300 mV s^{-1}. The pulse amplitude was 10 mV and current samples were taken immediately before each pulse and 5.5 ms after the application of the pulse, the pulse spacing being 12 ms. Although this voltammogram was obtained with an unusually high scan rate, the resolution is clearly very much improved. The m.s.p.s.a. potentiograms show a resolution which compares well with that of d.p.a.s.v., although when several scans are added to increase the sensitivity the resolution is decreased.

No comparisons between the sensitivities of the techniques are possible on the basis of these experiments, because with the sample considered the noise levels were very low and the signal-to-noise ratio difficult to estimate.

The support of the Danish Natural Science Research Council (grant number 511-8032) is gratefully acknowledged.

REFERENCES

1 S. P. Perone, Anal. Chem., 43 (1971) 1288.
2 Q. V. Thomas, L. Kryger and S. P. Perone, Anal. Chem., 48 (1976) 761.
3 T. Anfält and M. Strandberg, Anal. Chim. Acta, 103 (1978) 379.
4 P. E. Sturrock, R. D. Clemensen and A. C. Hayman, IVth International Conference on Computers in Chemical Research and Education, Novosibirsk, USSR, 1978, in press.
5 L. Kryger and D. Jagner, Anal. Chim. Acta, 78 (1975) 251.
6 J. Mortensen, E. Ouziel, H. J. Skov and L. Kryger, Anal. Chim. Acta, 112 (1979) 297.
7 S. P. Perone, D. O. Jones and W. F. Gutknecht, Anal. Chem., 41 (1969) 1154.
8 L. L. Miaw, P. A. Boudreau, M. A. Pichler and S. P. Perone, Anal. Chem., 50 (1978) 1988.
9 J. W. Dillard, J. A. Turner and R. A. Osteryoung, Anal. Chem., 49 (1977) 1246.
10 D. Jagner, Anal. Chem., 50 (1978) 1924.
11 D. Jagner and A. Graneli, Anal. Chim. Acta, 83 (1976) 19.
12 H. J. Skov, L. Kryger and D. Jagner, Anal. Chem., 48, (1976) 933.
13 S. P. Perone and J. F. Eagleston, J. Chem. Educ., 48 (1971) 317.
14 T. Anfält and D. Jagner, Anal. Chem., 47 (1975) 759.
15 L. Kryger, D. Jagner and H. J. Skov, Anal. Chim. Acta, 78 (1975) 241.
16 J. Nyborg, B. Nielsen and T. Bogh, RC4000-ALPHA Cross-assembler, University of Aarhus, 1977.
17 W. T. de Vries, J. Electroanal. Chem., 9 (1965) 448.

Analytica Chimica Acta, 122 (1980) 193—200
Computer Techniques and Optimization

ON-LINE COMPUTERS IN CLASSICAL CHEMICAL ANALYSIS†

M. BOS

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

(Received 9th July 1979)

SUMMARY

Developments in digital electronics enable classical chemical analysis to regain some of the terrain lost to instrumental techniques. In chemical analysis, computerization can provide higher precision, higher speed and lower costs. The value of interactive systems in routine work is emphasized.

The exponential way in which digital hardware is developing has considerable impact on the concepts of laboratory instrumentation and automation. Automation in the laboratory not only relieves the experimenter from the tedious acquisition of data, lengthy calculations, drawing of graphs, etc., but in many cases enables higher precision in measurements and the optimization of procedures with regard to speed. In the field of classical chemical analysis, laboratory equipment is traditionally simple and relatively cheap, but the procedures are laborious and complex. This is in contrast to instrumental methods where the situation is somewhat reversed. Not surprisingly, computers were first applied to instrumental analysis, primarily to optimize the use of expensive equipment. Increasing labour costs and decreasing costs of computers and secondary memory now favour the introduction of computers in classical chemical analysis. Of course, the extent to which automation of an analytical method is justified depends strongly on the type of method and the goal of the analysis. In practice, it varies from situations in which a computer is used as a back-up peripheral for a specific analytical instrument to a fully computerized method incorporated into a laboratory computer network. It is as well to realize that the natural goal of instrument designers, i.e. completely automatic operation, often still produces some human tasks that are very unattractive.

Motives for computerization, technical possibilities and some human aspects are considered in the following sections.

---

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia in September 1979.

MOTIVES FOR LABORATORY COMPUTERIZATION

There is a variety of situations in which automation can resolve problems in the laboratory. The reasons for automation can be roughly divided into two groups: technical motives and reduction of human labour costs. The frequently-encountered technical problems that can be solved by automation are: (i) information is needed very quickly after the experiment or measurement is finished; (ii) an instrument has a high rate of data output; (iii) there is a complicated relationship between the information wanted and the quantities measured; (iv) the experiment has a complicated measurement protocol; (v) the experiment requires control information gained during the course of the experiment. Applications of on-line computers in the laboratory with emphasis on reduction of human labour mostly concern data acquisition, report generation and simple pre-programmed routine control actions.

COMPUTERIZATION OF UNIT OPERATIONS AND MEASUREMENTS IN CLASSICAL CHEMICAL ANALYSIS

Classical chemical analytical methods can be characterized as based on a chemical or electrochemical reaction. In this field, two fundamentally different categories can be recognized; (i) absolute stoichiometric methods in which all of the compound to be determined is converted, e.g., gravimetry, titrimetry and coulometry; and (ii) dynamic methods in which a reaction rate or a stationary state is correlated with the amount of the compound to be determined via a calibration procedure, e.g. polarography, voltammetry and potentiometry. Generally, the dynamic methods can be readily adapted for continuous operation. However, they are intrinsically less accurate than stoichiometric methods. Methods of both categories consist of a number of unit operations and measurements which differ widely in the ease with which they can be automated. A survey of the situation is given in Table 1, which is based on liquid samples. The handling of solid or gaseous samples is very much more difficult to automate. Other than current and voltage control, the unit operations can normally be controlled by reading and operating on/off switches. This task can easily be accomplished by any on-line computer, which offers much greater flexibility than discrete logic circuitry. Programming these functions at bit level can best be performed in assembly language.

Operations to position samples require mechanical or pneumatic gear that can be expensive, needs maintenance, and is prone to failure. They should be avoided as much as possible.

Traditionally, analog devices have been used to control current and voltage. Modern high-speed digital hardware enables on-line computers equipped with analog-to-digital converters (ADC) and digital-to-analog converters (DAC) to perform these tasks. Again the greater flexibility is an enormous advantage. Very complex voltage or current versus time patterns can be generated in this way. Use of the measured data in generating the control action can provide optimization with regard to speed and accuracy.

The computerization of a measurement generally boils down to the use of

TABLE 1

Ease of automation of unit operations and measurements in chemical analysis

Unit operation	Ease of automation	Measurement	Ease of automation
Sampling	—	Weight	+
Sample transfer	—	Volume	—
Reagent addition	+	Time	+
Stirring	+	Temperature	+
Heating	+	Light intensity	+
Distilling	—	Current	+
Cleaning	—	Voltage	+
Filtering	—		
Voltage control	+		
Current control	+		

a transducer that converts the quantity of interest to a voltage and digitalization of this voltage by an ADC. An exception is the measurement of time. Generally, on-line computers incorporate a high-precision oscillator, and counting its pulses provides a time base. It is noteworthy that many manufacturers of measuring equipment now provide digital outputs and inputs on their instruments, thus facilitating computer control of the instruments. Even standardization is appearing, as is testified by the acknowledgement of the IEEE 488 bus.

DATA TREATMENT AND TYPICAL MATHEMATICAL PROCEDURES

A large part of the software written for laboratory automation is specific to its application, and deals with complex calculations converting raw data to meaningful results with the use of algorithms based on the theory of the phenomenon studied with the computerized equipment. However, several processes recur in a variety of applications: signal averaging, digital filtering, direct digital control, and the display of results. This warrants some attention to these processes.

Signal averaging

For measurements which can be repeated, multiple data sets are collected in successive runs and finally averaged point by point. This improves the signal-to-noise ratio by a factor equal to the square root of the number of times the measurement is repeated. This technique is known as ensemble averaging.

Digital filtering

Noise of a specific frequency can be removed from signals by filtering. Analog filters have been used to this end but tend to be difficult and expen-

sive and cannot be easily adjusted to changing experimental conditions. A digital filter that is used very often is the one presented in the classic paper of Savitzky and Golay [1]

$$\bar{Y}_j = \left(\sum_{i=-m}^{i=m} C_i Y_{j+i} \right) \Big/ N$$

where Y represents the data samples, N is the number of data points ($N = 2m + 1$, called the span of the filter), $\bar{Y}$ represents the filtered points, and C the filter coefficients, the values of which determine the operation of the filter.

Another type of filtering uses the Fast Fourier Transform [2]. It has been shown by Hayes et al. [3] that for electroanalytical data this method is somewhat faster than the Savitzky and Golay method, and the nature of the filter action can be interpreted more conveniently. The technique can also be used for interpolation of sampled electrochemical data [4].

Direct digital control

In titrimetric and electroanalytical methods there exist a number of control problems. Probably the best known are the control of the electrode potential (or sometimes current) in various electroanalytical methods, and the control of titrant addition (or generation for coulometric titrations) in set-point titrations. Much of the commercial gear available incorporates only some proportional control action for these processes. From industrial process control, it is known that much can be gained by the use of more elaborate controllers: i.e., with proportional, differential and integrating action. Eelderink et al. [5] have shown this for pH set-point coulometric titrations, which they were able to carry out in 12 s. These kinds of control actions can also be generated by software, and this is in fact common practice in industrial process control. Pomernacki and Harrar [6] showed the success of this approach for the control of electrode potential in controlled-potential electrolysis.

Display of results

The development of visual display units and plotters has opened up the way to almost perfect overviews of the results of automated measurements and calculations. Associated software (graphics) can be obtained commercially from the various computer manufacturers and software houses.Though a rather specialized field, this is of great importance as it can often be the determining factor in the success of a computerized method. The interactive use of graphics is very efficient in dealing with complex data as it combines the outstanding abilities of man and machine.

Although the need for visual inspection of the data, even for routine work with a well-tested method, has been recognized [7], the use of graphics is still not widespread in computerized titrimetric and electroanalytical methods. This is probably a result of the still relatively high cost of display hardware

and software. Developments in this area on the personal computer market show the use of cheap video graphic or alphanumeric displays. Skov et al. [8] chose this approach for a 256 × 192-point graphic display for analytical applications.

Typical mathematical procedures

The possibilities offered by computers in the field of complex calculations have led to better use of data. On the one hand, there is a tendency to use all available data to obtain statistically improved results. On the other hand, work is being done to minimize measurements while simultaneously improving their accuracy in order to cut down analysis times [9]. Both approaches require rather complex calculations which are only feasible when done by computer.

In attempts to use all available data to extract information, two fundamentally different categories of methods can be distinguished. The first category includes the multiparametric curve-fitting methods; these methods require a theoretical equation relating the measurement variables to a number of parameters that describe the experiment. Starting with initial estimates, the parameters are adjusted in the calculations until the theoretically calculated values provide a best fit to the experimental results according to a least-squares criterion. The general curve-fitting program written by Meites and Meites [10] has been used in numerous applications of titrimetry and electroanalysis, e.g. in the evaluation of weak base titrations [11] and in the processing of overlapping polarographic curves [12]. The general program can be easily adapted to new problems by inserting the equation relating the experimental data to the parameters required. Not so easy to modify but much faster in execution are the multiparametric curve-fitting programs based on the mathematics given by Wentworth [13]. Recent applications can be found in titrimetry [14, 15] and polarography [16, 17].

In addition to the multiparametric curve-fitting methods, the first category includes methods which use linearization of potentiometric titration plots to obtain equivalence volumes [18—21]. Modified Gran functions play an important part in these linearization procedures.

In the second category, a statistical rather than a physico-chemical model is used to correlate the data measured and the information wanted. Early applications, mainly based on the Linear Learning Machine of Nilsson [22], were introduced in mass and infrared spectrometry by Kowalski, Jurs and Isenhour [23, 24]. The first application of these pattern recognition methods in classical analysis was reported by Sybrandt and Perone [25]; this concerned the deconvolution of severely overlapping peaks obtained by polarography at a hanging mercury drop electrode. The applications remained restricted to qualitative results until Wold et al. [26] recently extended their SIMCA method [27] to quantitative work, and Bos and Jasink developed systems for the quantitative evaluation of data from anodic stripping voltammetry [28] and potentiometric acid—base titrations [29].

AUTOMATED ANALYTICAL METHODS AND TECHNICAL STAFF

In the development of automated analytical methods, the tendency to allocate as many functions as possible to the equipment often leads to a residual set of unpleasant tasks for the analytical technician. Especially in classical chemical analysis, there is a fair chance that he will be left with sample preparations and book-keeping. Moreover, loss of control over the analysis can sometimes be frustrating, especially in cases where there is no interaction during the course of the analysis. A third cause of problems is equipment that hurries the operator.

In order to ensure smooth operation of an analytical laboratory and the wellbeing of the technical staff, attention should be paid not only to the design of automated equipment but also to the design of meaningful jobs for its users. Rijnsdorp [30] gives a number of rules for the allocation of functions between "man" and "machine" to create meaningful jobs that encompass a variety of tasks that exercise the operators' skills. He stresses the importance of close cooperation between those responsible for the technical subsystem (the technical engineers), those responsible for the social subsystem (the social engineers) and the (future) users of the technical subsystems.

CONCLUSIONS

It is clear that the computerization of classical chemical analysis is only at its beginning. The trend is certainly towards decentralized use of computing power by means of the application of microcomputers. Except perhaps for a few exotic electroanalytical techniques, the requirements with regard to the speed and the amount of memory of computers for classical analytical methods are rather modest. Hardware costs will therefore present no major problem. Much attention should be paid to software development and the design of methods that can be operated in continuous-flow systems in order to ensure the continuing use of classical analytical methods in an economic way. The great diversity of methods in this field points to multi-purpose laboratory equipment with flexible software that allows for interactive optimization of analytical procedures. However, there will remain a need for specialized equipment, especially for monitors in industrial processes and environmental control. Naturally the demands with regard to the computer part of such equipment are different, the emphasis being on reliability and speed of operation. Integration of analytical and process control operations will certainly lead to improved operation of chemical processes with less waste products.

No doubt some of the newer analog signal-processing devices will be important building blocks in this field. Worth mentioning are the tapped analog delay line, the potentialities of which have been described by Horlick [31], and the so-called analog microcomputer, a LSI combination of a fast ADC, a simple microprocessor, some EPROM and RAM memory, and a fast DAC.

As for the software, the availability of computer power (16-bit micro) in the laboratory will allow sophisticated data-processing methods. A very promising technique is the Kalman filter [32], the use of which enables real-time digital optimization of analytical information. Also pattern recognition types of data processing should find increased application, especially in cases where the relation between the information wanted and the data measured is complex. It is important to note that although the learning methods require lengthy calculations in training, the final recognition calculations are simple and very fast.

Finally it should be stressed that real achievement in the automation of classical chemical analysis will depend on integration of knowledge from analytical chemistry, electronics, and computer science. University courses in analytical chemistry, traditionally of a multidisciplinary character, are best suited to accomplish this integration.

The author wishes to thank Mrs. B. Verbeeten-van Hettema for preparing the manuscript.

REFERENCES

1 A. Savitzky and M. J. E. Golay, Anal. Chem., 36 (1964) 1627.
2 J. W. Cooley and J. W. Tukey, Math. Comput., 19 (1965) 297.
3 J. W. Hayes, D. E. Glover, D. E. Smith and M. W. Overton, Anal. Chem., 45 (1973) 277.
4 R. J. O'Halloran and D. E. Smith, Anal. Chem., 50 (1978) 1391.
5 G. H. B. Eelderink, H. B. Verbruggen, F. A. Jutte, W. J. van Oort and B. Griepink, Z. Anal. Chem., 280 (1976) 273.
6 C. L. Pomernacki and J. E. Harrar, Anal. Chem., 47 (1975) 1894.
7 L. Kryger, D. Jagner and H. J. Skov, Anal. Chim. Acta, 78 (1975) 241.
8 H. J. Skov, L. Kryger and D. Jagner, Anal. Chem., 48 (1976) 933.
9 A. Olin and B. Wallen, Talanta, 24 (1977) 303.
10 T. Meites and L. Meites, Talanta, 19 (1972) 1131.
11 D. M. Barry and L. Meites, Anal. Chim. Acta, 68 (1974) 435.
12 W. F. Gutknecht and S. P. Perone, Anal. Chem., 42 (1970) 906.
13 W. E. Wentworth, J. Chem. Educ., 42 (1965) 96.
14 M. Bos, Anal. Chim. Acta, 90 (1977) 61.
15 L. M. Schwartz and R. I. Gelb, Anal. Chem., 50 (1978) 1571.
16 M. Bos, Anal. Chim. Acta, 103 (1978) 367.
17 M. Bos, Anal. Chim. Acta, 81 (1976) 21.
18 C. McCallum and D. Midgley, Anal. Chim. Acta, 78 (1975) 171.
19 V. Eliu-Ceauşescu and D. Ceauşescu, Z. Anal. Chem., 291 (1978) 42.
20 D. Midgley and C. McCallum, Talanta, 21 (1974) 723.
21 D. Midgley and C. McCallum, Z. Anal. Chem., 290 (1978) 230.
22 N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
23 B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 1945.
24 P. C. Jurs, B. R. Kowalski and T. L. Isenhour, Anal. Chem., 41 (1969) 21.
25 L. B. Sybrandt and S. P. Perone, Anal. Chem., 44 (1972) 2331.
26 C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.
27 S. Wold and M. Sjöström, ACS Symp. Ser., 52 (1977) 243.

28 M. Bos and G. Jasink, Anal. Chim. Acta, 103 (1978) 151.
29 M. Bos, Anal. Chim. Acta, 112 (1979) 65.
30 J. E. Rijnsdorp, Elektrotech. Maschinenbau, 96 (1979) 251.
31 G. Horlick, Anal. Chem., 48 (1976) 783A.
32 P. F. Seeling and H. N. Blount, Anal. Chem., 48 (1976) 252.

Analytica Chimica Acta, 122 (1980) 201—213
Computer Techniques and Optimization

PRINCIPLES AND PROBLEMS OF COMPUTER-BASED INSTRUMENTS AND NETWORKS IN ANALYTICAL CHEMISTRY†

H. C. SMIT

Laboratory for Analytical Chemistry, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam (The Netherlands)

(Received 5th November 1979)

SUMMARY

Some aspects of computerization in analytical chemistry are considered. Particular attention is given to the possibilities of the micro—mini—mainframe-computer network already developed. Two examples of applications, an automatic titrator and correlation high-performance liquid chromatography (h.p.l.c.), are described in order to provide heuristic insight and ready comprehension of the methods. Correlation chromatography is emphasized and a striking example of decreasing the detection limit in h.p.l.c. is given.

The aim of analytical chemistry is to acquire information concerning the quantitative or qualitative composition of mixtures, and this should be accomplished in an optimal way; that is, optimal with regard to accuracy, cost, time, etc., as determined by the user of the information. The proved and potential possibilities of computers in reaching a better optimum and in increasing the amount and value of the information is inarguable. This applies to the syntactic the practical and in some respects even to the semantic level of information.

In analytical chemistry a distinction can be made between computers dedicated to one task and computers for general-purpose applications; laboratory automation can be based on either type. The microprocessor appears to provide an intermediate solution. A central (mini)computer can be connected with microprocessors, acting as intelligent terminals, capable of carrying out not too complicated (pre-)calculations and possibly closed-loop control. Of course, the design can be directed to fixed applications for each terminal, dedicated to one analytical task. In this laboratory, however, a flexible system was needed, with each terminal suitable for divergent analytical applications. Computer-based analytical methods have been developed, where such a flexible system would have been a great help and would have reduced the development time considerably.

Two types of computer-based analytical methods can be distinguished:

†This paper was presented at the International Conference on Computer-based Analytical Chemistry, Portorož, Yugoslavia, in September 1979.

computerized "traditional" analytical methods and new, computer-dependent methods. Both categories have been studied in this laboratory and an example of each will be described.

COMPUTERIZED TRADITIONAL METHODS

Automatic titrator

The automatic determination of a titration curve involves simulation of the work of an analyst; new, computer-based possibilities are not created. The important step in designing such instrumentation is to make an inventory of the actions of a real analyst, determining a titration curve with, for instance, a pH meter as the basic instrument. Filling burets, etc., is very easily mechanized and automation is restricted to the determination of the titration curve with the following actions: (1) estimation of the amount of titrant to be added as a function of the slope of the already determined part of the titration curve; (2) adding the titrant; (3) waiting for equilibrium; (4) recording the pH as a function of the amount of titrant added. Establishing the slope of a curve means determination of the derivative of the pH—titrant volume curve. Equilibrium implies that the pH remains constant, i.e., that the time derivative of the pH is zero.

These analytical actions can be translated into computer actions: (a) calculation of the slope $\Delta\text{pH}/\Delta V$ of the titration curve from recently determined points; (b) introduction of $(\Delta\text{pH}/\Delta V)$ in an appropriate mathematical expression to calculate the new amount ΔV_n of titrant to be added (ΔV_n must decrease with increasing $(\Delta\text{pH}/\Delta V)$ and, conversely, increase with decreasing $(\Delta\text{pH}/\Delta V)$); (c) transformation of ΔV_n in a number of steps (pulses) for a stepping motor controlled buret; (d) generation of a voltage proportional to the total amount of titrant added $V + \Delta V_n$, controlling the x-deflection of an x—y recorder and (e) generation of a print-command pulse if the approximation of the time derivative $(\Delta\text{pH}/\Delta t)$ enters a small zero-centered interval. Some parameters like maximum step size, maximum total amount of titrant to be added, etc. have to be introduced for optimal performance in the determination of different types of titration curves.

All the actions mentioned have been performed by various electronic instruments based on analog techniques. However, calculation of the necessary dynamic range, especially in the case of the derivative of the titration curve, shows that a microprocessor-based solution is by far preferable to an analog solution. A microprocessor-based instrument was developed here in 1975. The necessary parameters can be introduced by means of thumb wheel switches. Some characteristic results are shown in Fig. 1. Experience with the automatic titrator, together with a more detailed description, will be published separately.

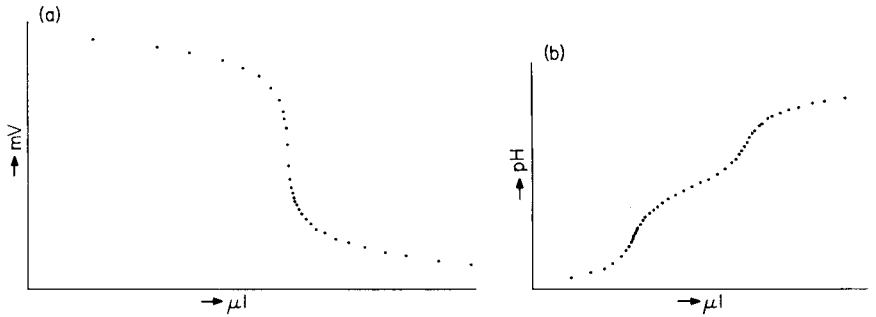


Fig. 1. Automatically determined titration curves for (a) 0.02 M $AgNO_3$ with 0.1 M KCl, and (b) 1.65×10^{-4} M H_3PO_4 with 0.1 M NaOH.

COMPUTER-DEPENDENT METHODS

Correlation chromatography

Correlation chromatography can be considered as an example of a computer application of the second category. i.e., an analytical technique which is impossible without the computer. Analytical methods are generally based on the limited data- and signal-handling capacity of relatively simple instruments and on the limited human capabilities in these respects. Thus it should prove rewarding to explore the possibilities of modifying well-known analytical methods or developing new techniques by utilizing the calculating and data-handling capacities of computers. The aim is to increase the analytical power of the method, by obtaining more information.

Correlation chromatography is an example of this approach and some details of this technique will be given without extensive theoretical considerations. More information has been given in references [1—5]. In conventional chromatography the sample is injected as a pulse and the response of the chromatographic system, including detector and recorder, is the chromatogram. In correlation chromatography the injection is not a pulse and the injection system is modified. The columns and the detector are the same in both systems.

Figure 2 shows a simplified diagram of a correlation chromatograph, in this case a h.p.l.c. system. The input of the column is connected in turn with the eluent and with a sample reservoir by means of electrically switched valves. The switching of the input is controlled by a spectral pattern, a so-called pseudo random binary sequence (PRBS). A PRBS is a kind of binary noise pattern. Binary noise is a noise with only two amplitude levels, $+a$ and $-a$ (Fig. 3). Nevertheless, pure binary noise is a stochastic signal, because it cannot be predicted which of the two levels, $+a$ or $-a$, will be present at a certain time, although on average each of the two levels will occur 50% of the time, i.e. each level has a probability of 0.5. Generally, a binary noise is created artificially by a generator controlled by an internal clock. The clock-period determines the minimum time during which one of the two states will exist; a shorter duration is impossible.

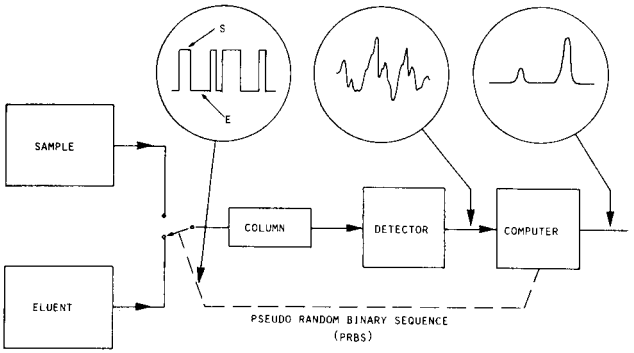


Fig. 2. Basic diagram of a correlation chromatograph.

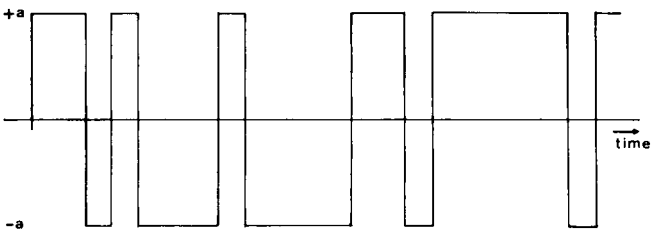


Fig. 3. Binary noise.

A pseudo-random signal has the remarkable property that it is not really random. After a certain time, or better, a certain number of clock-periods (a sequence), exactly the same pattern is repeated. It follows that a PRBS is fully described by the clock-period, the sequence length and the two amplitude levels [6]. When a PRBS is used to control the input valves of a correlation chromatograph, the two levels correspond simply to supplying eluent and supplying sample, respectively. The concentration of the components in the sample determines the amplitude of the response of the chromatographic system. Of course, this response also depends on the kind of input pattern and is the sum of the responses of the chromatograph on the injection of the single components in the sample, injected in the same pattern.

Figure 4 shows the detector output of a chromatograph as a result of a semi-continuous PRBS injection. The injected sample contains only two components (phenol and dimethylphenol) at a rather high concentration. All peaks are confluent and, at first sight, separation is out of the question. But, given this output together with complete knowledge of the PRBS input pattern, it is possible mathematically to calculate the "straightforward" chromatogram such as would be obtained by conventional pulse-shaped injection.

The procedure used to construct this chromatogram, or better correlogram, is to cross-correlate the detector output with a PRBS, similar to the injection PRBS mentioned, but with levels +1 and −1. Equation (1) shows the mathematical procedure resulting in the cross-covariance function

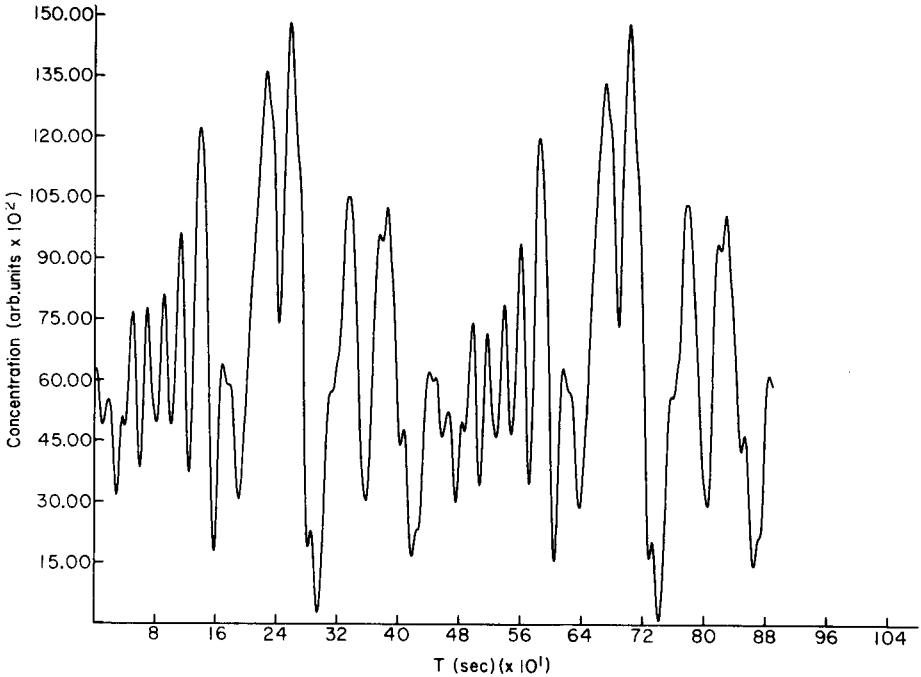


Fig. 4. Detector output after PRBS injection of sample with two components.

$$R_{xy}(\tau) = \lim_{T\to\infty}(1/T) \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t-\tau)\, y(t)\, \mathrm{d}t \qquad (1)$$

Cross-correlation means determining the average product of one signal $y(t)$, i.e. the detector output, with the time (τ)-delayed version $x(t-\tau)$ of another signal $x(t)$, i.e. the PRBS. Each value of the delay τ supplies one point of the resulting correlogram $R_{xy}(\tau)$ after the point-by-point products of both signals have been averaged. To make it plausible, without many mathematical derivations, that $R_{xy}(\tau)$ is identical to a conventional chromatogram, the autocovariance function $R_{xx}(\tau)$ of a binary noise has to be considered

$$R_{xx}(\tau) = \lim_{T\to\infty}(1/T) \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t-\tau)\, x(t)\, \mathrm{d}t \qquad (2)$$

Autocorrelation means multiplying a signal $x(t)$ with a time τ-delayed version $x(t-\tau)$ of the same signal $x(t)$. The average product is determined for every τ. Suppose that the time delay τ exceeds the clock-period Δt_c of a binary noise $x(t)$ with the levels $+1$ and -1, each with probability 0.5. Multiplying $x(t)$ and $x(t-\tau)$ point by point, the probability of obtaining a product $+1$ or a product -1 is the same. The average product in this case is zero. If $\tau < \Delta t_c$, e.g. $\tau = k\Delta t_c$ $(0 < k < 1)$, then the average product will be $(1-k)$.

When the range of τ is extended to negative values, $R_{xx}(\tau) = 0$ for every τ, except for the range $-\Delta t_c < \tau < \Delta t_c$ where $R_{xx}(\tau)$ is a triangle with top amplitude 1 (Fig. 5).

In the following argument it is assumed that a sample with one component is injected on a column according to a binary noise pattern, and that each input pattern in that particular column is delayed only with a retention time T_R; there are no changes in peak shape, no dispersion and no non-linearities. An injected pulse, as is usual in conventional chromatography, results in an identical output pulse after a time T_R. Of course, these assumptions are not realistic, but are useful in explaining the results of a correlation procedure. The binary noise pattern also remains unchanged; the pattern is delayed only by a time T_R.

Execution of the correlation procedure between output and input in this case means determining the average product of the time τ-delayed version $x(t-\tau)$ of the binary noise $x(t)$ and the time T_R-delayed version $x(t-T_R)$ of $x(t)$. The resulting $R_{xy}(\tau)$ will be a triangle, not with the top at $\tau = 0$, but at $\tau = T_R$. If the clock-period $t_c \ll T_R$, then the basewidth $2\Delta t_c$ is negligible with respect to T_R and the triangle can be considered as a narrow pulse. The height of the pulse depends on the concentration of the component in the input sample.

Each real chromatogram with several peaks with increasing peak width, obtained by a pulse-shaped injection, can be considered to be built up from infinitely narrow pulses, each with a certain amplitude and time delay. PRBS injection of this mixture of components and cross-correlation of the output with the PRBS results in a correlogram, identical to the chromatogram mentioned, because the procedure described for one pulse (point) is valid for every point in the chromatogram. Parts of the baseline, where no peak is present, converge to zero.

Application of a truly stochastic signal as input test signal for a system with the use of finite averaging time, introduces an uncertainty in the statistical determination of the system parameters. The pseudo-random property of a PRBS, which in fact is a strict function of time, and the use of a whole

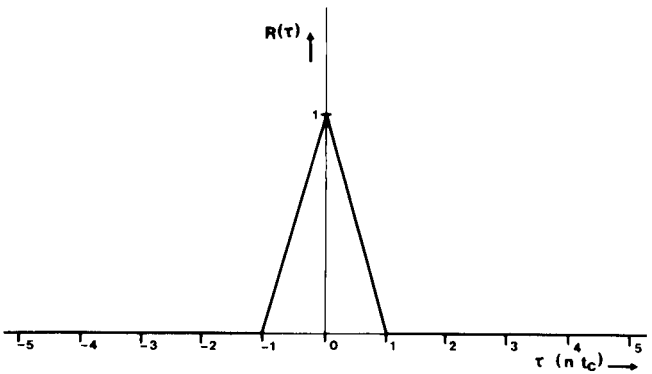


Fig. 5. Auto-covariance function of binary noise.

number of sequences as averaging time reduces this random error to zero.

The most important advantage of this complicated way of obtaining a chromatogram is that the detection limit can be lowered very efficiently. The noise of the system (detector noise, electronic noise, etc.) is not correlated with the PRBS, and the average of the products for τ will converge to zero with increasing averaging time. In other words, the cross-covariance function of the noise and the PRBS converges to zero. Theoretically, it is possible to achieve an infinitely low detection limit without preconcentration, but at the cost of time and sample volume. In practice, however, there are limitations. Apart from time and sample volume limitations, the most difficult restriction is imposed by the non-stationarity of the chromatographic system.

If the retention indices vary during the correlation procedure, e.g. because of flow variations, then the consequence is not only peak broadening, but also increasing baseline noise of the correlogram. This is caused by the fact that the average products are no longer converging to zero.

Complete theoretical derivations of the influence of non-stationarity have been given [7, 8]. An important result is that if the non-stationarity is known, later correction becomes possible. In Fig. 6, the influence of non-stationarity is shown. A cross-correlogram is determined by means of a simulated chromatographic column which, in the case of conventional chromatography, would produce a peak shifting in time together with increasing peak width. No system (detector) noise is present. Peak broadening and a fluctuating baseline in the correlogram is the result (Fig. 6, dotted line). Later correction by a rather complicated computer program gives a

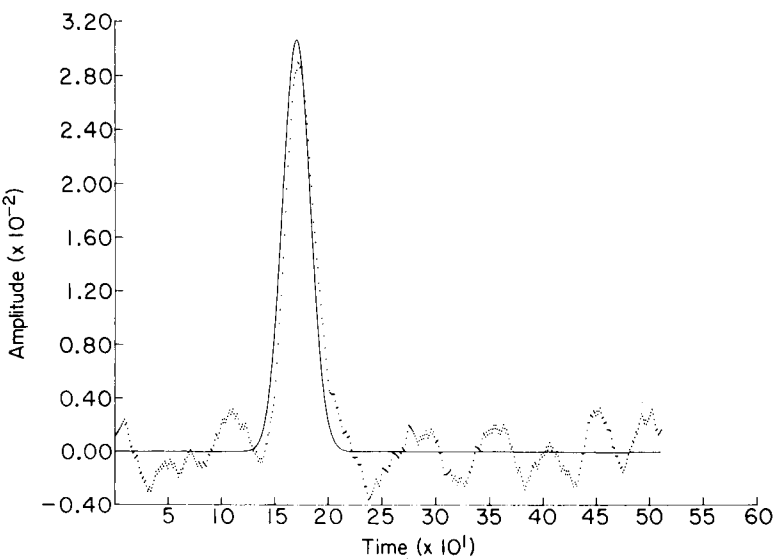


Fig. 6. Cross-correlogram of non-stationary system, corrected and uncorrected.

correct peak, as shown. This method would probably give another reduction of the detection limit, but only a computer simulation has so far been realized.

The practical realization of a correlation chromatograph implies two rather different aspects. The first problem is the construction of an injection system; the second is the design of an instrument capable of generating suitable PRBS sequences and of running the correlation procedure (a correlator). After experience had been gained with an experimental injection system and off-line signal handling, an injection system for correlation h.p.l.c. and a microprocessor-based correlator were developed in this laboratory. Both will be described in detail in separate publications, but a short description of the principles and problems will be given here.

A simplified diagram of the sample holders and injection system is shown in Fig. 7. Of course, the real system is much more complicated; normal injection is possible, a simple membrane pump for filling the system is present, etc. The system has to fulfil far-reaching demands concerning pressure (up to 500 atm) and corrosion resistance. The switching valves are positioned directly after the pump and not after the reservoirs because corrosion-resistant valves suitable for high pressure are difficult to obtain. The use of two reservoirs separated by pistons resistant to almost every solvent, makes it possible to use a non-corrosive liquid for the pump and the valves. The position of the piston (i.e. the amount of sample remaining) was determined by building in a small magnet and using a magnetic field detector with a Hall element and LED's as indicators.

Concerning the correlator, all demands and much more can be fulfilled by a microprocessor. The microprocessor-based instrument developed is combined with a video display (an ordinary TV set). The design philosophy was to develop an instrument usable for conventional chromatographers unskilled in the use of computers, microprocessors, data handling, mathematical background, etc. The only information which has to be introduced into the instrument, by means of a keyboard and three push buttons, is the time duration of the complete chromatogram, the standard deviation of the unretarded peak and, optionally, the number of sequences, i.e. the total correlation time. Using these parameters, the instrument determines the optimal

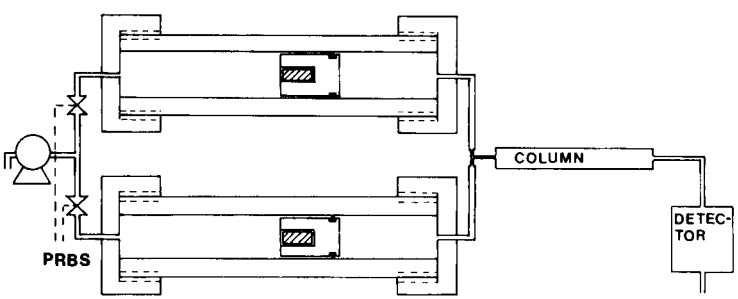


Fig. 7. Sample holders and injection system (simplified).

value of parameters like clock-frequency, sequence length, sample frequency of the A/D converter, cut-off frequency of the anti-aliasing filter, etc.

It is possible to monitor the detector output on the display, the correlation procedure and the intermediate results after complete sequences, as well as simultaneously to punch the detector output on paper tape. After the correlation procedure is finished, the correlogram appears on the display. A joystick control is present to control a flashing plus sign, a cursor, which can be shifted to each point of the correlogram. The amplitude and time of the point of the correlogram indicated by the cursor is displayed on the screen. The cursor can also be used to place marks. These marks allow selection of points from the baseline indicating the limits of part of the correlogram, e.g. the limits of a peak or the limits of a peakless part of the baseline. These indicated points can be used to enlarge part of the correlogram, for baseline "drift" correction, for determining the standard deviation σ_n of the noise in the peakless part of the baseline, and for area determination by peak integration with defined integration limits. The standard deviation of the area σ_I is calculated, and also displayed, by using the relation between σ_n and σ_I derived previously [9]. Of course, the final correlogram can be punched on paper tape, so that a plot can be made later. Figure 8 shows pictures directly photographed from the screen.

Figure 9 shows a plot of a conventional chromatogram representing the h.p.l.c. separation of twelve different chlorinated phenols. The corresponding h.p.l.c. correlogram is shown in Fig. 10. The most striking property is the increase in sensitivity given by the correlogram. Making allowance for real injection volumes, the estimated detection limit is decreased by a factor of ~40 in 2.25 h.

Apart from some minor differences in resolution (peaks 10 and 11) caused by a slight modification in separation conditions (temperature), the changes in peaks 2 and 12 should be noted. Peak 12 is probably a decomposition product or some isomer, but peak 2 appears in the wrong place. Some pollution, probably originating from the eluent, has given rise to a peak somewhere

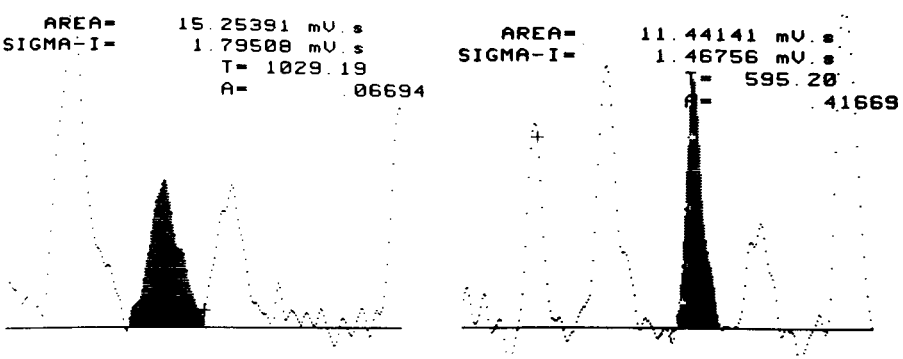


Fig. 8. Enlarged parts of a correlogram as displayed on the screen. The calculated area is related to the shaded peak.

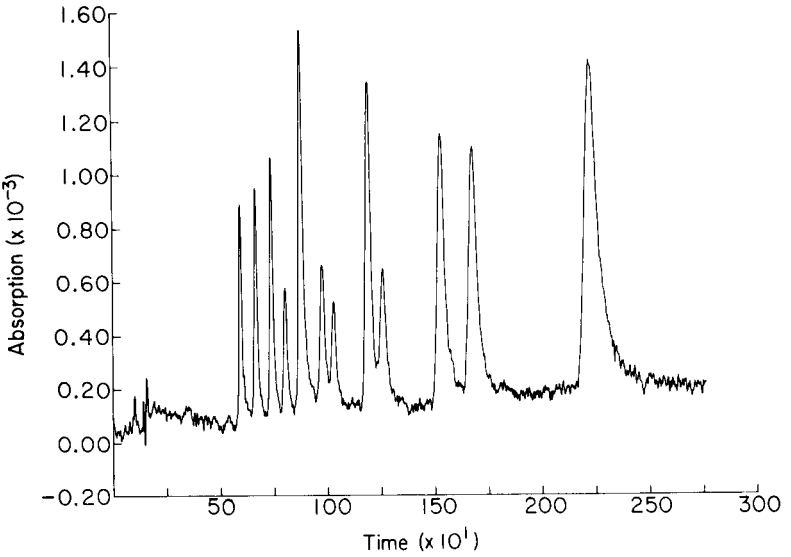


Fig. 9. Separation of 12 different chlorinated phenols by conventional h.p.l.c. (10 ppm).

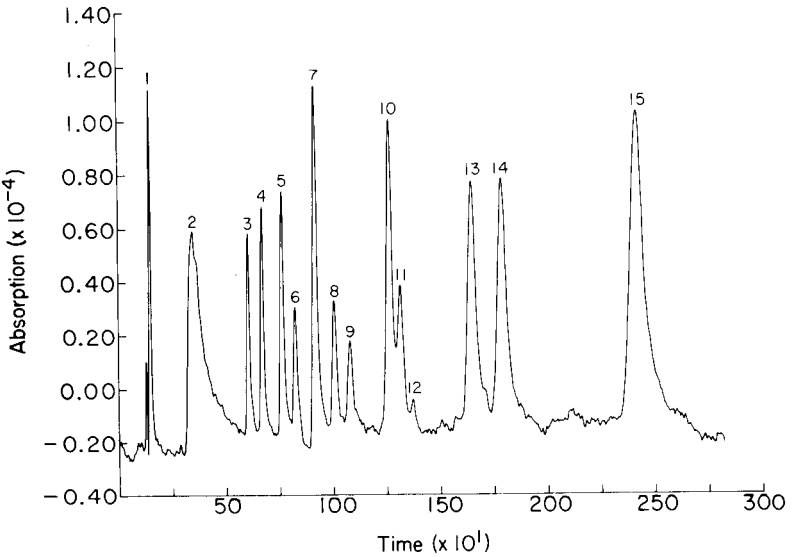


Fig. 10. Correlogram corresponding to Fig. 9 with slightly different separation conditions (0.2 ppm).

after the last "regular" peak. Probably the chosen duration of the chromatogram was too short, so that a peak was folded back in the correlogram, fortunately in an empty part; however, this explanation is rather speculative.

COMPUTER NETWORK

The design of a flexible general-purpose computer network, usable for analytical tasks such as the titrator and correlator described above, is now considered. The commercially available single microprocessor is a powerful instrument, but if it is to be used for very different applications, its limits will soon be reached, particularly with regard to memory capacity. A minicomputer coupled to one or more microprocessor systems is to be preferred. The general scheme was to use the minicomputer to develop software for the microprocessors and to use it as a library for the different microprocessor subroutines developed for analytical applications.

Cross-assemblers and the general advantages of a minicomputer operating system, e.g. editing possibilities, disc, etc., facilitate the development of the microprocessor software considerably. The microcomputer, loaded from the minicomputer, will perform the selected analytical tasks, including closed loop control of an analytical instrument. Minicomputer subroutines can be called and the results can be used by the microprocessor program. Extension of the tasks of a general-purpose system by memory-demanding and run time-demanding procedures such as pattern recognition, data-base management, etc., requires coupling of the network to a large mainframe computer, e.g. CDC, Cyber, etc.

The network developed on the basis of the concepts mentioned is shown in Fig. 11. The minicomputer (a Varian V76) can be considered as the heart of the network. The protocol processor, which is necessary for the (slow) conversation between the mini and the large mainframe university computer, is based on a hardware design; the task is performed by a microprocessor. In contrast to a software solution, one of the advantages is that the minicomputer can be replaced by another type without many troubles or costs.

Concerning the microcomputer system, a terminal mode program in ROM makes the microprocessor transparent to the micro terminal operator during its contact with the minicomputer. Transparent to both sides, the terminal input is transferred to the minicomputer and back after software data buffering. A very important additional task of this program is to recognize a

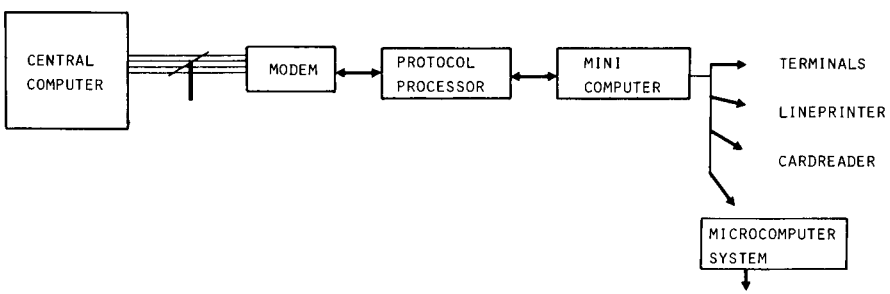


Fig. 11. Hardware network configuration.

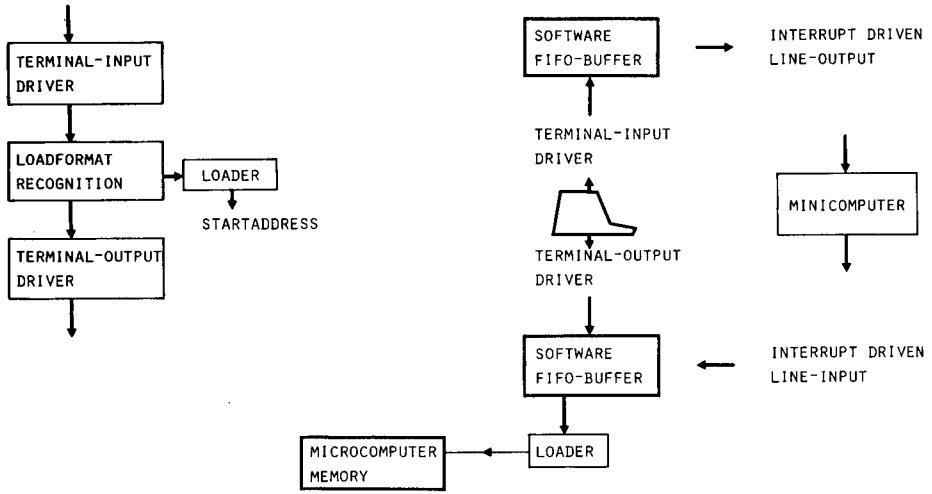


Fig. 12. Terminal-mode software and hardware configuration.

special load format from the minicomputer (Fig. 12). This load format does not appear on the terminal but goes directly in the memory of the microprocessor. This format provides a start address to direct the microprocessor to a chosen program.

In cooperation with a macro cross-assembler in the minicomputer, the combination has the following possibilities. First, the program development for the microprocessor can be carried out with the advantages of a larger operating system. Examples are file handling on disc, fast assembling and fast object loading in the microprocessor. Secondly, the macro cross-assembler allows programming in a pseudo-higher language, with self-defined macroprograms, i.e. assembler programs with their own instruction name. The microprocessor system is equipped with an 8-channel, 12-bit A/D converter and a number (24 per card) of parallel I/O gates, allowing closed loop applications.

In the future, an instrument bus controller will be connected to the I/O bus of the microcomputer. The microsystem can be loaded from the minicomputer with a kind of laboratory language interpreter, which makes it possible to program the system in BASIC-like manner. Optional total unique switching and measuring functions can be combined in an interpreter statement. A statement of the same kind can be used to order the microprocessor to make contact with the minicomputer, for instance to start a calculating program or exchange information.

The properties of this network will enable the analyst to automate an extensive number of analytical actions, including automatic titrations and correlation techniques. If a suitable interpreter-based language is developed, ad hoc automation of relatively simple procedures like titrations seems to be possible. The efficient program-developing properties facilitate automation

and optimization of more complicated procedures considerably. In that case, however, the use of assembler and self-defined macro programs will probably often be necessary to develop efficient programs which are not too time-consuming. This means that computer knowledge is indispensable. However, the execution of already developed programs requires no special training and allows the use of the network for routine analysis.

REFERENCES

1 H. C. Smit, Chromatographia, 3 (1970) 515.
2 R. Annino and L. E. Bullock, Anal. Chem., 45 (1973) 1221.
3 R. Annino, J. Chromatogr. Sci., 14 (1976) 765.
4 S. B. Philips and M. F. Burke, J. Chromatogr. Sci., 14 (1976) 495.
5 T. T. Lub, H. C. Smit and H. Poppe, J. Chromatogr., 149 (1978) 721.
6 A. C. Davies, IEEE Trans. Computers, C20 (1971) 270.
7 T. T. Lub and H. C. Smit, Anal. Chim. Acta, 112 (1979) 341.
8 M. Kaljurand and E. Küllik, Proc. of the 14th Int. Symp. on Advances in Chromatography, Lausanne, 1979, p. 173.
9 H. C. Smit and H. L. Walg, Chromatographia, 8 (1975) 311.

Analytica Chimica Acta, 122 (1980) 215–222
Computer Techniques and Optimization

CORRELATION OF OBJECTIVE CHEMICAL MEASUREMENTS AND SUBJECTIVE SENSORY EVALUATIONS. WINES OF *VITIS VINIFERA* VARIETY 'PINOT NOIR' FROM FRANCE AND THE UNITED STATES

WING-ON KWAN** and B. R. KOWALSKI*

Laboratory for Chemometrics, Chemistry Department, University of Washington, Seattle, WA 98195 (U.S.A.)

(Received 23rd October 1979)

SUMMARY

Forty wines of *Vitis Vinifera* var. 'Pinot Noir' were analyzed for their elemental and organic compositions by atomic emission spectrometry and glass-capillary gas chromatography, respectively. Their sensory quality was evaluated by a panel of judges. Average overall quality scores were used to rank these wines. Stepwise regression analysis and principal component factor analysis were used to investigate correlations between objective chemical measurements and subjective sensory evaluations. Compounds which were found to be related to overall quality of wine were identified.

Flavor research has benefited from advances in analytical instrumentation [1, 2]. Application of gas chromatography and mass spectrometry has allowed the separation and indentification of many flavor compounds in beverages [3–6] and foodstuffs [7, 8]. Multivariate statistical techniques are rapidly becoming essential tools in correlating subjective properties to objective properties of food products. This is manifested in the number of articles which have appeared in the literature, especially those associated with food industries [9–11]. Many research workers in areas which deal with the subjective–objective properties of consumer products, have pointed out that a combination of gas chromatography, mass spectrometry and computer data analysis, is becoming the most efficient system for their areas of research [12–17].

Wine is an interesting and challenging complex mixture of chemicals. Attempts to identify all essential flavor components in wine often yield little information about the organoleptic qualities of the individual components or about their influence on flavor acceptibility. Therefore, an approach is developed in this study which utilizes gas chromatographic patterns directly, temporarily bypassing the chemical identification of the components. Each sample is represented by an assembly of features. Each gas chromatographic peak is a feature and the normalized area under the peak is used as the magni-

**Present address: Carnation Research Laboratory, 8015 Van Nuys Boulevard, Van Nuys, CA 91412, U.S.A.

tude of the particular feature. The aroma-related features are sought by pattern recognition analysis. Once the chemical constituents that are most related to a sensory evaluation test have been identified by the pattern recognition and statistical methods applied to the features, only these key features need be identified by gas chromatography—mass spectrometry.

To understand the advantages and limitations of such an approach, the following characteristics are significant. First, pattern recognition analysis draws on the entire objective information that becomes available within the resolution limits of the particular analytical procedure. Secondly, components indicated to be most relevant to the flavor classification may simply signal the prominence of some flavor- and aroma-producing processes, but may or may not be significantly characteristic odorants by themselves.

In this study, elemental and organic compositions of 40 wines of *Vitis Vinifera* var. 'Pinot Noir' from France and the United States were correlated with their sensory evaluations. Components which were found to be most related to the overall quality and principal factors of the sensory evaluation scores were identified. The approach taken aims at displaying and summarizing a given set of data by means of numerical calculations. Results of this type can be classified as ad hoc relations between sensory and chemical data. Such relations are essential to the understanding of two other main categories of psychophysical relations between chemical and sensory measurements, namely "predictive" and "causative" relations.

EXPERIMENTAL

The forty 'Pinot Noir' wines from France and the United States used here have been listed by Kwan et al. [18]. There was a total of 154 chemical measurements, of which 137 were organic components [19], and 17 were elemental concentrations [18]. Fourteen sensory evaluation scores were obtained for each wine sample [20]. These included 13 individual evaluation parameters and the overall quality score. The 154 chemical measurements are called "features" and the overall quality score is regarded as a "property" of each wine. The importance of the 154 chemical measurements in correlation with the overall quality score was investigated by correlation and stepwise multiple regression analysis. Principal component factor analysis was also applied to both the sensory scores and chemical measurements. Detailed descriptions of these methods have been given by Harper et al. [21]. Features which showed high correlations with sensory properties were identified by gas chromatography—mass spectrometry (Energy Resources Co., Cambridge, MA).

RESULTS AND DISCUSSION

For the 40 wines, correlation coefficients were used to determine the importance of each of the 154 chemical measurements to the overall quality score of each wine. A list of chemical components which had high correlations with

the overall quality score is given in Table 1. 1-Hexanol, which had the highest correlation coefficient with the overall quality score, was also found in an earlier study [19] to be the most important chemical component that classified French and American 'Pinot Noirs' geographically. 2,4,5-Trimethyl-1,3 dioxolane also had a high variance weight for the same geographic classification [19]. But it was not chosen by the SELECT program because it provided information for classification very similar to that of 1-hexanol. Both of these organic compounds showed high discriminating power in geographic classification as well as high correlations to the overall quality scores. The elements, Sr, Ba, K, Ca and Mg, all showed negative correlations to the overall quality scores. They were also important chemical features used in geographic classification. Examination of the sensory evaluation results in Table 2 indicates that French 'Pinot Noirs' had overall quality scores higher than American 'Pinot Noirs'. The 40 wines could therefore be classified into two categories by either one of two criteria. The first criterion was the geographic origin which separated these 40 wines into French and American vintages. But the second criterion, which was the overall quality, gave similar classification. Therefore, those components which showed high correlation with overall quality, might have been selected because they fulfilled the criterion for geographic classification. Thus, it cannot be said that they are responsible for the difference in overall quality of these wines.

In order to eliminate the geographical factor, correlations of the chemical measurements with the overall quality scores were then performed on the French and American 'Pinot Noirs' separately. Important chemical measurements having high correlations are shown in Table 1. Two alcohols, namely 3-methyl-1-butanol and 1,3-dimethoxy-2-propanol, were found to correlate positively to the overall quality of American 'Pinot Noirs,' while pyruvic acid and phosphorus showed negative correlations. For the French 'Pinot Noirs',

TABLE 1

Correlations of chemical compositions to overall quality scores of wine

Origin	Organic	Elemental
American and French 'Pinot Noirs'	1-Hexanol (0.5850) 2,4,5-Trimethyl-1,3-dioxolane (0.5026)	Sr (−0.6914) Ba (−0.6122) K (−0.5812) Ca (−0.5593) Mg (−0.5012)
American 'Pinot Noirs'	3-Methyl-1-butanol (0.6064) Pyruvic acid (−0.5431) 1,3-Dimethoxy-2-propanol (0.4958)	P (−0.4466)
French 'Pinot Noirs'	1,1-Diethoxyethane (0.7657) Propylene glycol (0.5311) Cyclohexane (−0.5115)	Al (−0.6870) Cd (0.5378)

TABLE 2

Sensory evaluation scores of French and American 'Pinot Noirs' based on evaluations by seven judges

Quality	French 'Pinot Noir'			American 'Pinot Noir'		
	Range	Mean	S*	Range	Mean	S*
Clarity	0.0—1.0	0.9	0.3	0.3—1.0	0.9	0.2
Color	3.3—8.8	5.2	1.7	1.3—7.7	4.5	1.4
Aroma and bouquet intensity	3.4—8.3	5.6	1.3	2.7—5.8	4.5	0.9
Aroma and bouquet character	4.2—8.5	6.1	1.1	1.8—5.5	4.0	0.8
Undesirable odors	0.5—3.6	1.7	0.8	1.5—5.0	2.8	1.1
Acidity	4.8—6.8	5.6	0.6	4.0—6.7	5.2	0.6
Sugar	3.3—5.0	3.9	0.5	3.0—5.6	4.7	0.6
Body	4.3—7.3	5.4	0.7	3.0—6.3	4.5	0.7
Flavor intensity	4.5—7.5	6.0	0.7	3.8—6.0	4.9	0.5
Flavor character	3.5—8.3	5.8	1.2	2.3—5.6	4.2	0.9
Oakiness	1.8—5.0	3.8	1.0	1.5—5.0	3.5	0.9
Astringency	3.4—5.3	4.4	0.5	2.3—7.0	3.8	0.9
Undesirable taste and flavor	0.8—3.8	2.0	0.9	1.0—6.2	2.9	1.2
Overall quality	12.4—17.4	14.5	1.4	7.1—13.0	10.6	1.6

1,1-diethoxyethane and propylene glycol indicated positive correlations, while cyclohexane and aluminium correlated negatively to the overall quality. Cadmium was the only element that showed positive, rather than negative, correlation to the overall quality scores. Chemical components characteristic of the overall quality of these two subsets of wines were different from those of the entire set of 40 'Pinot Noirs'. When the 40 wine samples were analyzed simultaneously this result suggested that components related to overall quality were selected mainly because of their importance in geographic classification rather than their actual correlations to wine quality. In addition, the difference in components which correlated with the quality of these two types of wine, showed that 'Pinot Noirs' from different regions had their own characteristics in chemical compositions which correlated with their sensory properties.

Stepwise regression analysis

Stepwise regression analysis was then applied to these two subsets of data. This method searches for a fit between the dependent variable, which was the overall quality score, and 154 independent variables, which were the chemical measurements. Figure 1 shows the features chosen for the two subsets of wines and the fit multiple correlation coefficients. For American 'Pinot Noirs', 3-methyl-1-butanol was selected first for having the highest correlation to the overall quality scores and propanol was chosen next. The selection of propanol improved the fit correlation, but it was not the second highest correlation feature to the sensory scores. In both cases, the use of two variables gave reasonable fits. Testing for alternative variables that may

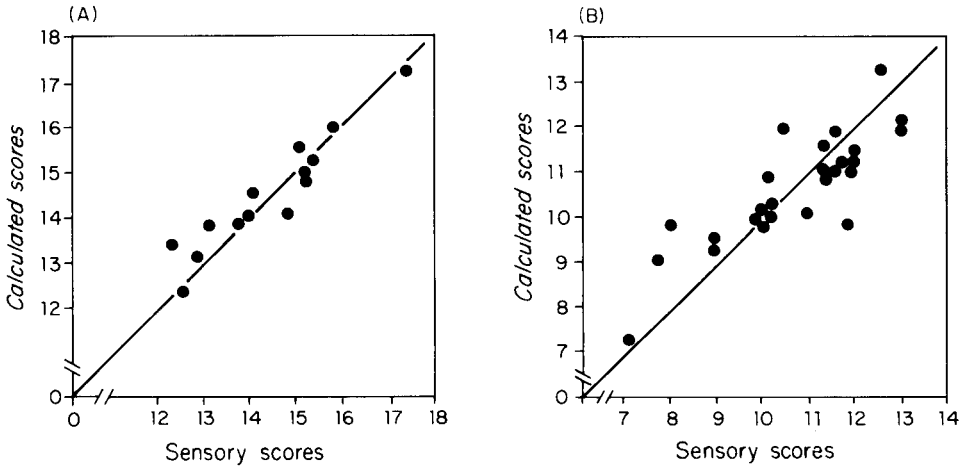


Fig. 1. Results of stepwise regression analysis of overall quality scores for (A) French 'Pinot Noirs' and (B) American 'Pinot Noirs' using 3-methyl-1-butanol and propanol. Correlation fit, 0.7521; standard error estimate, 1.11.

successfully replace those selected by statistical methods might also be instructive. If two chemical measurements are themselves highly correlated, and both correlate well to a sensory property, the stepwise regression method might pick one rather than the other by chance. The one being discarded might be equally as important in contributing to the actual sensory response. Therefore, equal emphasis should be placed on features with high correlations to the sensory property such as those in Table 1, as well as those which would improve mathematical fit as shown in Fig. 1.

Principal component factor analysis

The data set was then analyzed by principal component factor analysis. Three principal components each were extracted from the 13 sensory parameters and 154 chemical measurements respectively. Varimax rotations [21] of the principal factors were used to enhance extraction of key features from less significant ones in contribution to a particular factor. Lists of these features are shown in Table 3. In evaluating the 40 'Pinot Noirs,' the seven selected judges used three major factors. The first principal factor was composed of sensory parameters related to aroma and flavor. In the sensory evaluation study of these wines [20], the judges placed strong emphasis on characters of both flavor and bouquet, which were correlated with their respective intensities as well as body. The second principal factor for sensory parameters was a combination of undesirable flavor and odor, which would definitely contribute negatively to the quality of wine. The third principal factor was a combination of acidity and sugar. Results [20] showed that the judges placed less emphasis on these two factors in evaluating the overall quality of wine. Correlation of these three principal component factors and those of the chemical features was investigated. Although the first chemical factor, which consisted of γ-butyrolactone, diethyl succinate and 2-phenyl-

TABLE 3

Principal component factor analysis of chemical measurements and sensory evaluations of 40 'Pinot Noirs' from France and the United States

Principal component factor	Sensory evaluations	Chemical compositions
First	Flavor intensity	γ-Butyrolactone
	Body	Diethyl succinate
	Flavor character	2-Phenylethanol
	Aroma and bouquet — character	
	Aroma and bouquet — intensity	
Second	Undesirable taste and flavor	Succinic acid
	Undesirable odor	Propanol
Third	Acidity	Propionic acid
	Sugar	Glutaric acid

ethanol, was found to relate to the sensory factor which was made up of aroma and flavor parameters, no significant correlation was observed for other factors.

No chemical element included here was found to contribute significantly to these principal factors. Although some elements had been suggested to have certain effects on the sensory properties of wine [22], organic components were the dominant factors in determining the sensory quality of wine in this case. An element whose concentration varied with a certain property of wine, might not have any direct effect on this particular property. But it would appear high on the list of components correlated to this particular property. Using factor analysis eliminates to some extent this type of correlation which, though mathematically sound, may be enologically meaningless.

Results obtained so far demonstrate the ad hoc relations between chemical measurements and sensory evaluations for this set of wines. Most of the chemical components which showed high correlations to overall quality scores and principal factors of sensory evaluations in this study, had previously been isolated and identified in wines of many other varieties of *Vitis Vinifera* [23—28]. Because these chemical components are commonly present in most wines, they may be good candidates to serve as indicators for the general sensory quality of wine. Although these components may not be stimuli, the above results may have potential in establishing predictive relations between chemical and sensory measurements. Regression models might predict sensory scores for wines similar to those from which the predictive relations have been derived, though predictive co-variation may not exist for wines not included in the original tests. The more the chemical system is changed, the higher is the probability that compounds which are not stimuli will lose their predictive power. True stimuli compounds will be less

affected. Obviously, it is very much easier to duplicate the experimental conditions for objective measurements than for subjective sensory evaluations.

The features which show high correlations to sensory properties may or may not have any direct effect on the sensory quality. They may simply be the by-product of some aroma-producing processes. Investigation of the possible sources of these key components may provide information on their relations with aroma compounds, which are directly responsible for sensory response. Although the mechanisms of the production of some families of chemical compounds in wine are understood, the pathways of generation of most of the specific components identified in this study have not yet been fully investigated. These components may or may not be flavor and aroma compounds themselves. Nevertheless, attempts to unravel the production processes of these key elemental and organic species may provide valuable information leading eventually to a better understanding of the direct stimulus –response mechansim.

CONCLUSION

This study has demonstrated the feasibility of an approach for systematic investigation of three different levels of relations between chemical and sensory measurements. The use of pattern recognition techniques and advanced instrumentation has established ad hoc relations for this set of data. The key components selected are commonly present in other wines, and their use as indicators for sensory quality may further understanding of predictive relations. From the very large number of chemical components in such a complex mixture as wine, selection of key components serving as possible candidates for true stimulus—response mechanism studies should simplify the task. This cost-effective approach may be welcomed by flavor researchers.

Obviously, long-range application of findings such as those described here for wine improvement will require much more additional work. Sensory and consumer preference studies are needed to establish the optimum combination of flavor dimensions. These would be followed by experiments to establish if artificially modifying some statistically important objective properties would shift the flavor dimensions in the expected direction. Lastly, investigations would be needed on how to alter the objective properties by raw material and process changes. Major interdisciplinary efforts on the part of physiologists, chemists, psychophysicists and enologists are required. These efforts would be fully justified if a basic understanding of flavor as it relates to food acceptance could be achieved.

The authors express their gratitude to Dr. C. A. Sleicher for help in wine evaluations, and to Maynarhs daKoven for valuable observations.

REFERENCES

1 J. Ayres and W. R. E. Clark, Food Manuf., 1 (1975) 19.
2 A. C. Noble, Food Technol., 12 (1975) 56.
3 M. A. Gianturco, R. E. Biggers and B. H. Ridling, J. Agric. Food. Chem., 22 (1974) 759.
4 C. G. Tassan and G. F. Russel, J. Food Sci., 39 (1974) 64.
5 L. L. Young, R. E. Bargmann and J. J. Powers, J. Food Sci., 35 (1970) 219.
6 R. E. Biggers, J. J. Hilton and M. A. Gianturco, J. Chromatogr. Sci., 7 (1969) 453.
7 A. Dravnieks, H. G. Reilich, J. Whitfield and C. Watson, J. Food Sci., 38 (1973) 34.
8 T. Persson, E. von Sydow and C. Akesson, J. Food Sci., 38 (1973) 682.
9 D. R. Godwin, R. E. Bargmann and J. J. Powers, J. Food Sci., 43 (1978) 1229.
10 H. R. Moskowitz and C. D. Barbe, J. Food Sci., 41 (1976) 567.
11 T. Persson and E. von Sydow, J. Food Sci., 39 (1974) 537.
12 J. J. Powers and E. S. Keith, J. Food Sci., 33 (1968) 207.
13 T. H. Parliment and R. Scarpellino, J. Agric. Food Chem., 25 (1977) 97.
14 C. Merritt Jr., D. H. Robertson, J. F. Cavagnaro, R. A. Graham and T. L. Nichols, J. Agric. Food Chem., 22 (1974) 750.
15 J. J. Powers and M. C. Quinlan, J. Agric. Food Chem., 22 (1974) 744.
16 P. Issenberg, A. Kobayashi and T. J. Mysliwy, J. Agric. Food Chem., 17 (1969) 1377.
17 J. J. Powers, Food Technol., 22 (1968) 383.
18 W-O. Kwan, B. R. Kowalski and R. K. Skogerboe, J. Agric. Food Chem., 27 (1979) 1321.
19 W-O. Kwan and B. R. Kowalski, J. Agric. Food Chem., in press.
20 W-O. Kwan and B. R. Kowalski, J. Food Sci., 45 (1980) 213.
21 A. M. Harper, D. L. Duewer, B. R. Kowalski and J. L. Fasching, in B. R. Kowalski (Ed.), ACS Symposium Series No. 52, American Chemical Society, Washington, DC, 1977, p. 14.
22 M. A. Amerine, Composition of Wines. II. Inorganic Constituents, Advances in Food Research, Vol. VIII, Academic Press, New York, 1958.
23 S. S. Chaudhary, A. D. Webb and R. E. Kepner, Am. J. Enol. Vitic., 19 (1968) 6.
24 R. E. Kepner, A. D. Webb and L. Maggiora, Am. J. Enol. Vitic., 20 (1969) 25.
25 D. J. Stern, A. Lee, W. H. McFadden and K. L. Stevens, J. Agric. Food Chem., 15 (1967) 1100.
26 D. J. Stern, D. Guadagni and K. L. Stevens, Am. J. Enol. Vitic., 26 (1975) 208.
27 C. J. van Wyk, R. E. Kepner and A. D. Webb, J. Food Sci., 32 (1967) 664.
28 A. D. Webb, R. E. Kepner and L. Maggiora, Am. J. Enol. Vitic., 18 (1967) 190.

Analytica Chimica Acta, 122 (1980) 223—226
Computer Techniques and Optimization

Short Communication

SIMPLEX OPTIMIZATION OF THE SYNERGIC EXTRACTION OF A BIS-DIKETO COPPER(II) COMPLEX

ROBERT J. McDEVITT and BARBARA J. BARKER*

Department of Chemistry, Xavier University, Cincinnati, OH 45207 (U.S.A.)

(Received 7th November 1979)

Summary. Simplex optimization is used to maximize the synergic extraction of copper(II) acetylacetonate into cyclohexane through formation of the isoquinoline adduct. Concentrations of isoquinoline and acetylacetone and pH were optimized for 1×10^{-4} M copper(II) solutions. Distribution coefficients of 0.25 were obtained above pH 7.9 for acetylacetone and isoquinoline concentrations above 3.0×10^{-4} M and 2.0×10^{-4} M, respectively.

Efficient experimental design provides answers to three problems [1]: the dependence of measured response on certain factors, the best-fit equation for such a dependence, and the optimal levels of the important factors. The present study was designed to establish, by applying simplex optimization, the optimal levels of the important factors in the synergic extraction [2] of copper(II) from a buffered aqueous phase into cyclohexane using acetylacetone (acac) and isoquinoline.

Theory

Simplex optimization, which evolved in the late 1950's and early 1960's, is a method for rapidly determining the region of optimum response of a chemical system by using multivariate techniques [3—8]. The simplex is defined by $n + 1$ points, where n is the number of variables in the system, and is made to progress through the factor space under consideration (the domain of the variables) by simple rules. By following the rules of operation, the region of optimum response is recognized by the circling of the simplex about the vertex most closely associated with the optimum.

The synergic effects obtainable with mixed-ligand systems in liquid—liquid distribution of metal ions are well established. Among many other extraction studies, Irving and Al-Niaimi [9] demonstrated the synergic enhancement of the extraction of copper(II) from acetate buffers into solutions of acetylacetone and 4-methylpyridine (Mepy) in benzene. They attributed the enhancement to the formation of the readily-extracted species $Cu(acac)_2$, Mepy. Similar investigations were made by using quinoline and isoquinoline [2]. As is usual in such studies, enhancement of distribution coefficients depended on the concentrations of acetylacetone and of the organic base and on the pH of the aqueous phase.

In complex systems such as these, the simplex optimization technique provides a valuable approach. The optimum levels of the significant experimental factors can be determined even without detailed knowledge of the formation constants and partition coefficients of the various species in solution.

Experimental

Reagents. Stock solutions of 0.0025 M copper(II) nitrate trihydrate in deionized water, 0.0025 M acetylacetone in cyclohexane, and 0.0025 M isoquinoline in cyclohexane were used. Cyclohexane was purified by distilling through a fractionating column packed with glass wool, the middle fraction (b.p. 80—82°C, lit. 81°C) being collected. Acetylacetone and isoquinoline were also distilled; the fractions boiling, respectively, at 138—139.5°C (lit. 139°C) and at 241—244°C (lit. 243°C) were used.

Stock buffer solutions were prepared [10] from solutions of 0.1 M potassium dihydrogenphosphate, 0.025 M sodium tetraborate, 0.1 M sodium hydroxide and 0.1 M hydrochloric acid. Distilled, deionized water was used throughout. Oxalyldihydrazide was prepared from diethyl oxalate and hydrazine hydrate [11] and recrystallized from boiling, distilled water.

Apparatus. An insulated oil bath, equipped with an appropriate heater-stirrer unit, cooling unit, stirring motor, thermoregulator, and thermometer, maintained the temperature of 25.00 ±0.02°C. A Beckman DB spectrophotometer was also used.

Procedures. The factors varied were the initial concentrations of acetylacetone and isoquinoline in the cyclohexane phase and the pH of the aqueous phase. The step size chosen for pH in the simplex scheme was 0.1 units. Fifty one buffer solutions covering the pH range 5.8—10.8 in increments of 0.1 unit were prepared; there was a change from phosphate to borax buffers at pH 8.1.

The initial concentration of the copper(II) in each experiment was 1.0×10^{-4} M; 1 ml of stock solution was diluted to 25 ml with the necessary buffer. The step size chosen for both the acetylacetone and the isoquinoline concentrations was 1.0×10^{-4} M.

Equal volumes of the copper(II) in the buffer solution and the acetylacetone—isoquinoline solutions in cyclohexane were shaken in 250-ml glass-stoppered flasks for 15 min at 25 ±0.02°C. After 5 min at rest in the bath to allow phase separation, a 10-ml aliquot of the aqueous phase was removed for determination of the copper(II) by the oxalyldihydrazide spectrophotometric method [11], using a conventional calibration graph.

Results and discussion

The simplex optimization can be summarized as follows [5]. The first step is to define and quantify the response to be optimized. Next, the factors to be varied are chosen, and step sizes, i.e., increments by which each factor is to be increased or decreased, are selected. Then the boundary

conditions of the factor space, i.e., the limits which cannot or should not be crossed during the search, are defined. The fifth step is to locate and construct the initial simplex.

The initial point may be taken as standard conditions for the factors if the system under consideration has previously established values for the factors, or an arbitrary point in factor space may be chosen. After the initial point has been chosen, the remaining n points of the initial simplex are defined relative to this point. Optimization then proceeds according to the rules of simplex methodology [3]. In the region of the optimum the simplex circles its vertex which lies closest to the optimum. This point then can be used as the initial point for a simplex of smaller step size in order to locate the optimum more precisely.

From the copper(II) concentrations determined, the distribution coefficient, K_D, was calculated by assuming that the difference between $[Cu(II)]_{in}$, the initial concentration of copper(II), and $[Cu(II)]_{aq}$, the final concentration of copper(II) in the aqueous phase, was equal to $[Cu(II)]_{org}$, the concentration of copper(II) in the organic phase, i.e. $K_D = [Cu(II)]_{org}/[Cu(II)]_{aq} = ([Cu(II)]_{in} - [Cu(II)]_{aq})/[Cu(II)]_{aq}$.

The initial values of the concentrations of acetylacetone and isoquinoline were chosen to correspond to the stoichiometric amounts needed to form the 1:2:1 copper—acetylacetone—isoquinoline complex [2]. The choice of initial pH was arbitrary. After the initial simplex had been constructed, the movement of the simplex followed the usual rules. The factor levels for the new vertex were calculated with a BASIC computer program SIMP 01 on a PDP 1/11 time-sharing computer. The program selected the vertex with the least desirable response from the four vertices which constituted the current simplex and then computed the values of the factors for a new vertex.

In Table 1, the results of the simplex optimization scheme used are summarized. A total of fifteen experimental observations was made, including one re-observation of a persistent vertex (vertex 9) and one observation of an arbitrary point in the factor space. The simplex moved along well and ultimately reached a plane of level response above a pH of 7.9 with initial concentrations of acetylacetone at approximately 3.0×10^{-4} M and isoquinoline at 2.0×10^{-4} M. This plane had a distribution coefficient $K_D = 0.25$. This level of response is similar to that obtained by Irving and Al-Niaimi [2] who used equilibration for 1 h with benzene.

The persistent retainment of vertex 9 necessitated re-observation of that point. The value of $K_D = 0.25$ was accepted as satisfactory because it conformed to the general trend of the data which showed a level response at higher pH values. The shift from a NaOH—KH_2PO_4 to an HCl—$Na_2B_4O_7$ buffer seems to have made no difference, probably because it came on the plateau of level response.

Antagonistic effects, which were attributed to complexes of the type CuB_j^{2+} (j = 1,2,3...), were observed by Irving and Al-Niaimi [9]. This type of response was not found here, but the lack of movement of the simplex

TABLE 1

Results of simplex optimization for the extraction of $Cu(acac)_2(iQ)$ between aqueous phases and cyclohexane[a]

Vertex	Vertices retained	[acac] ($\times 10^4$ M)	[iQ] ($\times 10^4$ M)	pH	K_D
1		2.0	1.0	7.5	0.00
2		2.0	1.0	7.7	0.10
3		2.9	1.0	7.6	0.02
4		2.3	1.8	7.6	0.14
5	2,3,4	2.6	1.4	7.8	0.14
6	2,4,5	2.0	1.6	7.7	0.02
7	4,5,6	2.6	1.2	7.7	0.15
8	4,5,7	3.0	1.9	7.7	0.25
9	5,7,8	3.2	2.0	7.9	0.46
10	7,8,9	3.2	2.0	7.7	0.09
11	8,9,10	3.9	2.1	7.9	0.25
12	8,9,11	3.5	2.0	8.0	0.25
13	9,11,12	4.0	2.1	8.2	0.25
9	Re-observation	3.2	2.0	7.9	0.25
14	Arbitrary point	4.0	3.1	10.0	0.25

[a]acac = acetylacetone; iQ = isoquinoline.

toward higher concentrations of isoquinoline was observed. Several factors may be responsible for the differences in the two sets of results.

The simplex technique of multivariate analysis proved to be quick and effective in determining the optimum levels of pH and reagent concentrations needed to maximize the synergic extraction of copper(II) into cyclohexane. Analysis of the response of the system indicates that the extraction depends mainly on the concentration of the organic base once the optimum levels of pH and acetylacetone concentration have been reached.

Studies of different types of nitrogen bases on the response would be of value in elucidating the nature of the interactions between the ligands and the metal ions.

REFERENCES

1 A. S. Olansky, L. R. Parker, Jr., S. L. Morgan and S. N. Deming, Anal. Chim. Acta, 95 (1977) 107.
2 H. M. N. H. Irving and N. S. Al-Niaimi, J. Inorg. Nucl. Chem., 27 (1965) 1671.
3 S. N. Deming and S. L. Morgan, Anal. Chem., 45 (1973) 278A.
4 G. S. G. Beveridge and R. S. Schechter, Optimization: Theory and Practice, McGraw-Hill, New York, 1970.
5 D. E. Long, Anal. Chim. Acta, 46 (1969) 193.
6 J. A. Nelder and R. Mead, Comput. J., 7 (1965) 308.
7 W. Spendley, G. R. Hext and F. R. Himsworth, Technometrics, 4 (1962) 441.
8 G. E. P. Box, Appl. Stat., 6 (1957) 81.
9 H. M. N. H. Irving and N. S. Al-Niaimi, J. Inorg. Nucl. Chem., 27 (1965) 717.
10 Handbook of Chemistry and Physics, CRC Press, Cleveland, OH, 57th edn., 1976—77, p. D-134.
11 G. Gran, Anal. Chim. Acta, 14 (1956) 150.

ACA announcements

ANNOUNCEMENTS OF MEETINGS

VTH INTERNATIONAL CONFERENCE ON COMPUTER IN CHEMICAL RESEARCH AND EDUCATION (VICCCRE)

VICCCRE will be held in Toyohashi, Japan, October 15–17, 1980, as a post congress symposium of 7th International CODATA Conference which will be held in Kyoto, October 8–11, 1980.

The conference covers almost all aspects of computer application in chemical research and education. There will be plenary lectures presented by the invited scientists and technical papers or short communications.

The title of the lecture together with an abstract of not more than 250 words should be submitted not later than August 1, 1980, to Prof. S. Sasaki, School of Materials Science, Toyohashi University of Technology, Tempaku, Toyohashi, Japan 440.

EUCHEM – CONFERENCE: Die Darstellung logischer Strukturen in der Chemie durch Modelle und die Lösung chemischer Probleme mittels Computern, October, 27–31, 1980, Endorf bei Rosenheim/Bayern, F.R.G.

The Scientific Committee (Chairman: Prof. Dr. I. Ugi) of the above-mentioned EUCHEM Conference have arranged the following plenary lectures to emphasise the theme of the conference:
J. Brandt and J. Bauer: Die systematische Darstellung und Voraussage chemischer Reaktionen
F. Choplin and G. Kaufmann: Computer Assisted Synthesis and Strategy in Organo Phosphorus Chemistry
R. Fugmann: Zur topologischen Darstellung von Begriffsbeziehungen
J. Gasteiger: Syntheseplanung mittels EROS
H. Gelernter: Making SYNCHEM useful and usable
Z. Hippe: Chosen Problems of Designing of Self-Adjusting System for the Discovery of Organic Syntheses
G. Derflinger and H. Keller: Neue mathematische Aspekte der Chiralität
B. Kowalski: Chemometrics and Pattern Recognition
M.F. Lynch: Representation and Searching of Generic Chemical Structure Using a Formal Grammar-Based Approach
G. Moreau: A New Molecular Descriptor: The Atocorrelation of the Topological Structure. Applications
P.J. Plath: Graphentheoretische Methoden in der Chemie
W. Schubert: Syntheseplanung etc. mittels ASSOR
I.K. Ugi et al: Die deduktive Lösung chemischer Probleme mittels Computern.

Prepared short discussion remarks relevant to the themes of the lectures are invited. For further details regarding preliminary registration and submission of discussion remarks please contact GDCh-Geschäftsstelle, P.O. Box 90 04 40, D-6000 Frankfurt/M 90, F.R.G.

CALENDAR OF FORTHCOMING MEETINGS

June 10–13, 1980 Ghent, Belgium	**3rd International Symposium on Quantitative Mass Spectrometry in Life Sciences** Contact: Professor A.P. De Leenheer, Laboratoria voor Medische Biochemie en Klinische Analyse, de Pintelaan 135, B-9000 Ghent, Belgium.
June 16–18, 1980 Milan, Italy	**7th International Symposium on Mass Spectrometry in Biochemistry, Medicine and Environmental Research** Contact: Dr. A. Frigerio, Istituto di Ricerche Farmacologiche "Mario Negri", Via Eritrea 62, 20157 Milan, Italy.
June 18–19, 1980 London, Great Britain	**Nuclear Magnetic Resonance Spectroscopy in Solids** Contact: The Executive Secretary, The Royal Society, 6 Carlton House Terrace, London SW1Y 5AG, Great Britain.

June 18–20, 1980 Brigham Young U, Provo, Utah, U.S.A.	**2nd Symposium on Environmental Analytical Chemistry** Contact: Delbert J. Eatough, 271 FB, Thermochemical Institute, Brigha Young U, Provo, Utah 84602, U.S.A.
June 23–27, 1980 Birmingham, Great Britain	**Eurochem 80** Contact: Andrew Dedman, Clapp & Polick Europe Ltd., 232 Acton Lan London W4 5DL, Great Britain.
June 26–29, 1980 Strasbourg, France	**International Symposium: Affinity Chromatography and Molecular Inte** Contact: Dr. J.M. Egly, Faculté de Médicine, Institut de Chimie Biologiq 11 rue Humann, 67085 Strasbourg Cédex, France.
June 30–July 4, 1980 Cannes, France	**13th International Symposium on Chromatography** Contact: GAMS, 88 Boulevard Malesherbes, 75008 Paris, France.
July 7–11, 1980 Brussels, Belgium	**2nd International Congress on Toxicology** Contact: Secretariat, SdR Assosiated, 16 Avenue des Abeilles, B–1050 Brussels, Belgium.
July 20–26, 1980 Lancaster, Great Britain	**SAC 80** Contact: The Secretary, Analytical Division, The Chemical Society, Burlington House, London W1V 0BN, Great Britain. (Further details published in Vol. 106, No. 2 and Vol. 113, No. 2)
Aug. 3–9, 1980 London, Great Britain	**Clinical Pharmacology and Therapeutics** Contact: Conference Associates, 34 Stanford Road, London W8 5PZ, Great Britain.
Aug. 4–8, 1980 Denver, Colo., U.S.A.	**Conference on Applications of X-Ray Analysis** Contact: Mrs. Mildred Cain, Denver Research Institute, University of Denver, Denver, Colo.,80208, U.S.A. Tel. 303/753-2141.
Aug. 4–9, 1980 Ottawa, Canada	**7th International Conference on Raman Spectroscopy** Contact: Mr. Ken Charbonneau, Conference Services, National Research Council of Canada, Ottawa, Ontario, Canada K1A OR6.
Aug. 17–23, 1980 Wolfeboro, N.H., U.S.A.	**Gordon Research Conference on Vibrational Spectroscopy** Contact: Dr. Erich Ipsen, Bell Laboratories, Holmdel, N.J. 07733, U.S.A.
Aug. 18–22, 1980 Brighton, Great Britain	**Micro 80** Contact: The Royal Microscopical Society, 37/38 St. Clements, Oxford OX4 1AJ, Great Britain.
Aug. 24–29, 1980 San Francisco, Calif., U.S.A.	**ACS 180th National Conference – 2nd Chemical Congress of the North Amer. Continent** Contact: A.T. Winstead, 1155 16th Street, N.W. Washington, D.C. 20036, U.S.
Aug. 24–31, 1980 Rzeszów, Poland	**2nd International Summer School on Data Processing in Chemistry DPC '80** Contact: Prof. Dr. Z. Hippe, Dept. of Physical Chemistry, Technical University, 35-959 Rzeszów, Poland. (Further details published in Vol. 122, No. 1)
Aug. 25–29, 1980 Prague, Czechoslovakia	**J. Heyrovský Memorial Congress on Polarography** Contact: Czechoslovak Academy of Sciences, J. Heyrovský Institute of Physical Chemistry and Electrochemistry, Vlasská 9, CS-118 40 Praha 1, Czechoslovakia (Further details published in Vol. 113, No. 2)
Aug. 25–30, 1980 Graz, Austria	**8th International Microchemical Symposium** Contact: Prof. Dr. A. Holasek, Institut für Medizinische Biochemie, Universität Graz, Harrachgasse 21, A-8010 Graz, Austria. Tel. (0 316) 32 5 32 or 76 5 91. (Further details published in Vol. 109, No. 1 and Vol. 110, No. 2).

VOL. 122, No. 2 ANALYTICA CHIMICA ACTA June 1, 1980

(Computer Techniques and Optimization, Vol. 4, No. 2)

CONTENTS

Printed in The Netherlands.