Vol. 103 No. 2 June 15, 1978
(Computer Techniques and Optimization, Vol. 2 No. 2)
ISSN 0378-4304

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

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

PUBLICATION SCHEDULE FOR 1978 (incorporating the section on Computer Techniques and Optimization).

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

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with every aspect of modern chemical analysis, both fundamental and applied. The section on *Computer Techniques and Optimization* is devoted to new developments in chemical analysis by the application of computer techniques and by interdisciplinary approaches, including statistics, systems theory and operation research. The section deals with the following topics: Computerized acquisition, processing and evaluation of data. Computerized methods for the interpretation of analytical data including chemometrics, cluster analysis, and pattern recognition. Storage and retrieval systems. Optimization procedures and their application. Automated analysis for industrial processes and quality control. Organizational problems.

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

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

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

Advertisements. Advertisement rates are available from the publisher.

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

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 8 volumes in 1978. The subscription for 1978 (Vols. 96–103) is Dfl. 1000.00 plus Dfl. 120.00 (postage) (Total approx. US $486.96). The subscription for the *Computer Techniques and Optimization* section only (Vol. 103) is Dfl. 125 plus Dfl. 15.00 (postage) (Total approx. US $60.87). Journals are sent automatically by air mail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–87) are available at Dfl. 115.00 (plus postage).

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

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

Analytica Chimica Acta, 103 (1978) 109—120
Computer Techniques and Optimization

APPLICATION OF A TEXT SEARCH SYSTEM BASED ON BOOLEAN STRATEGY TO MASS SPECTRAL DATA IDENTIFICATION

JAMES A. de HASETH**, HUGH B. WOODRUFF§, STEPHEN R. LOWRY‡, and THOMAS L. ISENHOUR*

The William R. Kenan, Jr. Laboratories of Chemistry 045A, University of North Carolina, Chapel Hill, North Carolina 27514 (U.S.A.)

(Received 14th February 1978)

SUMMARY

A general algorithm for text searching, operated on a tape-based minicomputer, has already been reported. This paper presents the application of the general text-searching algorithms to the *Registry of Mass Spectral Data* of 18,806 different entries. The text format allows multi-information input to be used to search the spectral library on the basis of data not necessarily extracted from mass spectra. Two library files have been generated; one is approximately half the size of the other, less important information having been deleted. The shorter library contains all 18,806 entries but enjoys much faster search times. Batch processing of searches is also possible. The text search is shown to be versatile in its operation, as the user can construct searches to be either broad or very selective, depending on the application. The search also has the capability to examine the data base internally and to check certain data for their validity.

Considerable interest has been generated in developing computer mass spectral search systems, especially since the development of g.c.-m.s. interfaces. This interest has produced many attempts at creating efficient and accurate computer algorithms for searching mass spectral data. The idea of a search to fulfil these needs was first proposed by Abrahamsson et al. [1], who thereafter developed a mass spectral search [2]. This early search relied on peak positions and intensity values of spectra to locate matches. Other approaches have been developed such as the use of discrepancy factors [3] and abbreviated spectral searching [4—6]. It was soon realized that peak positions in mass spectra provide very selective signatures, in which intensity information may be eliminated and the spectra reduced to peak—no-peak information [7—9]. In attempts to produce more stringent output, more information has been incorporated into search systems, such as molecular weight and spectral

**Present address: Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37916.
§Present address: Merck Sharp and Dohme, Research Laboratories R50A-102, P.O. Box 2000, Rahway, New Jersey 07065.
‡Present address: T. R. Evans Research Center, Diamond Shamrock Corporation, Painesville, Ohio 44077.

features [10—12]. Further investigations have led to the incorporation of pattern recognition techniques into a basic search, to determine better input parameters [13, 14]. A conversational search has also been implemented where the user interacts with the computer in order to produce a concise result [15—18]. One search has incorporated statistical hypothesis testing to improve the search performance [19]. Probability theory has been used to devise a search system that operates with probability-based matching [20]. Most searches have involved arithmetic testing criteria but the area of text-searching mass spectral data has been virtually left untouched. Text searching has the advantage that not only spectral information is searched, e.g. peak positions, but name, formula, molecular weight, elemental composition, structure, etc. may also be included. This added flexibility employs many of the useful aspects of arithmetic searching while having the ability to implement valid information that is usually omitted.

Text searching involves the direct comparison of all characters, alphabetical, numerical or punctuational, in one common mode. The input to the search is a profile which is compared to a library of known entries. During the search there are no numerical computations to assess a match or to calculate discrepancy or similarity indices. The input information is not abbreviated or used to compute a spectral "signature", nor are any match factors compiled. The input profile is compared directly with each entry in the library and only exact matches are output. The output is controlled by the input profile which permits the user to be as selective or broad as is required. The input profile is virtually unlimited in its flexibility, hence the search itself is very versatile.

THEORY OF OPERATION

In a text search system, the entire spectral file is converted into text or character format, i.e. all data, even data normally expressed as numeric or binary data, are converted to computer-readable characters. In the system described here, the data file is stored on magnetic tape. (The format compatible with this text search system requires that all spectral data be in ASCII characters.)

As all data fields, e.g. mass position and compound name, are in one common mode, any set of fields can be searched easily. Because the user can select the fields applicable, the search is very flexible. By the selection of input information, individual profiles can be constructed to accomplish both broad and very narrow searches. The use of Boolean logic allows profiles of any necessary complexity to be constructed. Other searches have, in effect, employed Boolean intersects by adding information to the library file, such as molecular weight and substructure units; however, the text search system presented here attempts to exploit more fully the use of the Boolean intersects and to improve search performance. When it is desirable to run more than a single search, the profiles can be batch-processed. As this system operates off-line, the batch-processing feature conserves time; hence, as more searches are run together, the time per search decreases.

The format used for the mass spectral text search system with a minicomputer is applicable to any chemical or literature data base. The search algorithms are general and only the library must conform to the system through its format. The same computer algorithms used in the mass spectral search system have proven useful with other data bases such as the Chemical Abstracts Chemical Condensates [21] and the ASTM Infrared Spectral Index [22] of 91,875 infrared spectra.

OPERATION

The data base used in this search system is the collection of 18,806 different compounds of the Registry of Mass Spectral Data (magnetic tape version, Wiley-Interscience, New York). The Registry of Mass Spectral Data, supplied by Professor F. W. McLafferty, includes an additional substructure bit code. These data are codes derived from the Wiswesser Line Notation for each compound, and although the code does not supply complete structure information, substructural characteristics may be ascertained. The original data set was written in a mixture of binary integer numeric format and EBCDIC character code. The entire file was converted to 8-bit ASCII character code to make the data base compatible with the algorithms previously generated for the above-mentioned search systems [21, 22].

All the data originally found in the McLafferty mass spectral data file were retained, except some of the molecular weight data. Four fields were omitted: the molecular weight rounded-off to the nearest integer; the integer molecular weight without isotopes; the exact molecular weight without isotopes, truncated to an integral value; and, the mantissa value of the molecular weight without isotopes, expressed as an integer. Additional fields were generated, giving the peak positions (*m/e*) of the five most intense peaks in order of descending intensity.

A single output citation is illustrated in Fig. 1 showing all 27 available search fields retained or generated from the original mass spectral data set. The serial number is followed by the compound name and the molecular formula. As can be seen, the molecular formula is expressed as a series of consecutive characters. The exact molecular weight follows the molecular formula which in turn is followed by the number of peaks and the *m/e* peak positions. Peak positions are expressed as integers when the *m/e* value is an integer; however, the mantissa is included to one decimal place if the *m/e* value is not an integer.

```
  151***TRIDEUTEROMETHYL ETHYL ETHER***CCC  DDD 0***63.076348***22***14 15
16 17 18 25 26 27 28 29 30 31 32 33 34 36 42 43 47 48 62 63 ***550 1165 614
355 3171 97 1035 1974 1067 3203 2880 32 453 1262 32 194 194 809 194 9999 93
8 226 ***X***48***29***18***30***27***0 1 2 01 02  ***X***X***X***X***RMU 5
      ***X***260***190***X***90***1***0 0***H5
```

Fig. 1. A sample output citation from the mass spectral library file. Each field is separated by three stars.

The peak intensity field follows with the largest peak normalized to 9999. The next six fields are the parent peak and five largest peak positions. After the fifth largest peak is the Wiswesser structural information code. This field does not contain the Wiswesser Line Notation (WLN) for the compound, but up to 27 codes derived from the WLN, which represent structural information about the compound. The remaining fields include instrumental and operational data except for the last which is the number of hydrogen atoms in the molecular formula. Hydrogen is not represented in the molecular formula field, as it would require too much storage space in the consecutive character format. An "X" in any field denotes that there is no information available for that field. For example, Fig. 1 has an X in the parent peak field, i.e. the parent peak was not discernible.

Any or all of the fields illustrated in Fig. 1 can be searched by the use of a search query or "profile." Profile fields are identified by a set of two-character mnemonics which are listed in Table 1. Each mnemonic corresponds to a single field in each citation. The order of the list in Table 1 corresponds to the fields presented in Fig. 1. Table 1 shows that many of the 27 fields are rather useless if employed as search input parameters; however, this does not preclude the data being useful as output.

If all 27 available data fields are used, a single search may take 12—15 min; thus, to speed up the search process, the number of data fields was reduced to 13. A sample citation of the shortened data set is shown in Fig. 2, and the fields that were retained are indicated in Table 1 by stars. The intensity field, which contains little searchable information, was not included on the short tape. The only fields which indirectly use intensity are the five largest peak positions, so that overall the text search observes only the presence or absence of peaks. A detailed examination of both text-search tape files

TABLE 1

Search system input mnemonics

MNEMONIC	PHRASE TYPE	MNEMONIC	PHRASE TYPE	MNEMONIC	PHRASE TYPE
SN*	SERIAL NUMBER	P2*	SECOND LARGEST PEAK	I1	INSTRUMENT
NA*	NAME	P3*	THIRD LARGEST PEAK	I2	INLET
FO*	FORMULA	P4*	FOURTH LARGEST PEAK	I3	ION SOURCE TEMPERATURE
WT*	EXACT MASS	P5*	FIFTH LARGEST PEAK	I4	INLET TEMPERATURE
NP*	NUMBER OF PEAKS	WN*	WISWESSER TERMS	ST	SPC TEMPERATURE
MP*	MASS POSITION	MS	METASTABLE MASS POSITIONS	EV	ELECTRON VOLTS
MI	MASS INTENSITY	RR	REARRANGEMENT MASS POSITIONS	IA	ION ACCELERATOR POTENTIAL
PP*	PARENT PEAK	DC	DOUBLY CHARGE MASS POSITIONS	PR	PRESSURE
P1*	LARGEST PEAK	IM	IMPURITY MASS POSITIONS	HN	NUMBER OF HYDROGEN ATOMS

```
151***TRIDEUTEROMETHYL ETHYL ETHER***CCC  DDD 0***63.076348***22***14 15
16 17 18 25 26 27 28 29 30 31 32 33 34 36 42 43 47 48 62 63 ***X***48***29*
**18***30***27***0 1 2 01 02
```

Fig. 2. The same citation as illustrated in Fig. 1 with many of the less useful input fields eliminated to shorten the library file.

showed that the reduced format is 47% shorter than the full format. Search times are further reduced by batch processing.

The equipment used for the previous search systems [21, 22] and this search system consists of a Raytheon 704 65,536 8-bit byte processor with a 1-μs cycle-time; two 25-ips 800-bpi magnetic tape drive units; a 500-cpm card reader; and a 300-lpm line printer. This equipment was not selected for any special abilities to effect a text search, but was used because of its availability.

SEARCH STRUCTURE

As described above, the input for the search is a profile. A few simple rules are adequate for constructing Boolean algebraic profiles from the defined fields (as represented by their mnemonics, see Table 1) and keystring labeling variables. The Boolean algebra operators applicable in the search are AND and OR. The AND operator means precisely that the result is true only when both inputs are true. The OR operator is the inclusive OR, and is true when either or both of the inputs are true. The NOT delimiter is not an operator but simply negates the inference of any input and must be used in conjunction with an operator.

An advantage of the text search is that information from measurements in addition to the mass spectrum can be conveniently used. Figure 3 gives an example of a profile that was constructed from several different types of spectra. The first two statements in Fig. 3, denoted by slashes, are comment statements and may be used to contain any information pertinent to the profile. (Comments are carried through the search and used as a header for the output.) The first keystring of the profile selects the P1 field which is the largest peak. A two-character user-assigned variable, A1, is used to identify the peak position at *m/e* 108. The rule for a keystring is always the same: the two-character mnemonic; a comma; a one- or two-character user-assigned variable; an equality sign; and the keystring information. In this example, three peaks were of sufficiently high intensity that any one of them might be the largest peak under slightly different operating conditions. After the three P1 keystrings there is a sub-predicate statement, denoted A. (This variable, also user-assigned, may consist of one or two characters.) The sub-predicate A is equivalent to keystrings A1 OR A2 OR A3.

The second largest, third largest and fourth largest peaks are entered into the profile in the same manner as the largest peak. The mass spectrum also contains thirteen significant peaks identified with the MP mnemonics. Here the sub-predicates E and F have all the keystrings AND'ed together, as all

```
/   UNKNOWN FROM SILVERSTEIN & BASSLER
/
P1,A1=108                                   MP,E1=108
P1,A2=43                                    MP,E2=43
P1,A3=91                                    MP,E3=91
A=A1:A2:A3                                  MP,E4=90
P2,B1=108                                   MP,E5=150
P2,B2=43                                    MP,E6=79
P2,B3=91                                    MP,E7=39
P2,B4=90                                    MP,E8=107
B=B1:B2:B3:B4                               MP,E9=51
P3,C1=108                                   MP,E0=77
P3,C2=43                                    E=E1&E2&E3&E4&E5&E6&E7&E8&E9&E0
P3,C3=91                                    MP,F1=50
P3,C4=90                                    MP,F2=65
P3,C5=150                                   MP,F3=89
P3,C6=79                                    F=F1&F2&F3
C=C1:C2:C3:C4:C5:C6                         FO,G1=*O
P4,D1=43                                    FO,G2=CCCCCCCCCC*
P4,D2=91                                    FO,G3=*N
P4,D3=90                                    FO,G4=*E
P4,D4=150                                   FO,G5=*G
P4,D5=79                                    FO,G6=*S
P4,D6=39                                    G=G1&-(G2:G3:G4:G5:G6)
P4,D7=107                                   HN,H=H10
P4,D8=51                                    **=A&B&C&D&E&F&G&H
P4,D9=77
D=D1:D2:D3:D4:D5:D6:D7:D8:D9
```

Fig. 3. A sample profile constructed from an unknown problem where the mass, i.r. and n.m.r. spectra were given. The Boolean algebra operators AND, OR and the delimiter NOT are symbolized by &, : and -, respectively. (This example was taken from R. M. Silverstein and G. C. Bassler, Spectrometric Identification of Organic Compounds, 2nd edn., Wiley, New York, 1967, p. 172.)

must be present, not just any one as in the largest peak sub-predicates A, B, C, and D. Finally, when the peak positions of the mass spectrum are examined, there is no evidence of the unknown compound containing nitrogen, chlorine or bromine. The lack of nitrogen is confirmed by the infrared spectrum which also indicates the absence of sulfur. The infrared spectrum, however, exhibits a very strong carbonyl band, therefore the compound must contain oxygen. Elemental presence or absence is searched by means of the FO mnemonic. All elements are indicated by their WLN symbols, e.g. E for bromine and G for chlorine. In this profile each single element is preceded by a star to indicate "truncation." Thus ★O means that at least one oxygen is present, because truncation allows any number of characters to precede the single O. (★O will also accept HO for holmium, etc., but these elements (Ho, Co, Mo, No, and Po) are very minor occurrences in the file.) The truncation star may be used at either end of a keystring, as in the keystring labeled G2 (FO field in Fig. 3). The n.m.r. spectrum is used to establish the presence of ten protons. The integrated peak areas indicate a factor of ten protons in the unknown. The spectral shifts also indicate a phenyl, a methylene and methyl group. The parent peak in the mass spectrum is at *m/e* 150, which is used as the approximate molecular weight. The source for this example supplied the information that *m/e* 150 is the parent peak; this information may not be generally available, but here it is included for illustrative purposes. Combining the molecular weight information with the requirement of 10 hydrogen atoms

and at least one oxygen means the molecule can have no more than 9 carbon atoms. Consequently, 10 or more carbon atoms cannot be present. The sub-predicate statement labeled G ties together all the elemental data. It reads: there is at least one oxygen atom present, but there cannot be more than 9 carbon atoms, or any nitrogen, bromine, chlorine, or sulfur. The final predicate statement, denoted **, completes the profile by stating that all the conditions imposed by each of the sub-predicate statements must be algebraically true, simultaneously.

The output that resulted from searching with this profile is presented in Fig. 4. As can be seen, only one citation of the 18,806 entries in the file was retrieved. The unknown is indeed benzyl acetate and this single result shows the specificity of this system. Furthermore, this result is not from the application of a refined profile; rather it was produced by a first attempt. It is interesting to note that the five largest peaks in the unknown mass spectrum are, in descending order: 108, 91, 43, 90, and 150; whereas, in the search system library file, they are: 108, 91, 90, 43, and 150. This illustrates the necessity of the multiple choice option for the largest peaks.

Although the rules may appear complex at first, a little practice and familiarity makes profile writing a simple, rapid procedure. Once the profile(s) has been written, three basic steps are needed to effect a search.

1. MASSSET. This program takes the input profiles and constructs a search structure from the keystrings. The search structure is an algebraic set designed for sequential and canonical searching of the data base. MASSSET replaces WARPSET in the other text-search systems [21, 22], as MASSSET will acommodate the 27 mnemonics needed for the mass spectral search, whereas WARPSET will handle only nine. Apart from this difference, the two programs are identical.

2. WARP-8. This program performs the actual multiprofile search. The program compares the input data with the spectral file, locates all matches, solves the Boolean logic defining the search strategy, and outputs the citations agreeing with any of the profiles temporarily onto a magnetic tape.

3. SORPRINT. As the output citations are on the magnetic tape in the order that they were found on the spectral file, they must be sorted with respect to the input profiles and printed. This program completes the search operation.

```
/   UNKNOWN FROM SILVERSTEIN & BASSLER
/

   4677***BENZYL ACETATE***CCCCCCCCC  00  ***150.068082***53***14 15 26 27 28
29 31 32 37 38 39 40 41 42 43 45 46 46.5 49 50 51 52 53 61 62 63 64 65 66 73 7
4 75 76 77 78 79 80 86 87 89 90 91 92 93 105 106 107 108 109 110 150 151 152 *
**30 270 30 130 100 70 20 70 40 110 651 40 130 80 3883 40 180 20 20 330 891 21
0 80 40 140 430 130 1091 60 20 90 60 70 1271 320 2292 210 30 30 1161 4224 5225
 400 20 310 60 1742 9999 781 50 3053 310 30 ***X***108***91***90***43***150***
V 0 R V0 01 1  ***X***X***X***X***X***X***252***202***X***X***X***X***H10
```

Fig. 4. The single output citation generated by the profile in Fig. 3 which shows the specificity that is possible with this search system.

RESULTS

As was seen with the profile in Fig. 3 and its search output in Fig. 4, the search algorithms can be very selective. Had the profile included less input information, there may have been more output citations. As all output citations must unequivocally meet all required input criteria, all are equally correct with respect to the profile. This artifact does not allow any ranking or indexing of the output data; therefore, absolutely correct answers are determined by stringent input requirements. These stringent input requirements may be incorporated into the profile when data are included from sources other than mass spectra.

In some search situations, it may be desirable to set broad input criteria. For example, it may be desirable to generate a subset of compounds for a specialized search system or other applications such as pattern recognition studies, i.e. if all compounds that contain a carbonyl group were needed for a particular study, a broad profile requesting only carbonyl presence in the WN field would suffice. Hence, profiles can be tailored to actual needs. The recall of the profile, i.e. the number of output citations, depends directly on the specificity of the profile. All output citations are correct or in absolute agreement with the profile; hence, the precision of the output, i.e. the number of correct citations in the output, is always 100%. Clearly, a profile has a finite probability of producing no citations if the profile is too stringent, or too many citations if the profile is too broad. Either case necessitates revising the profile and rerunning the search. This process, although time-consuming, is interactive and does produce refined results.

Figure 3 shows clearly that there is uncertainty in some of the fields, i.e. the largest peak fields. The profile did not require that fields P1, P2, P3 and P4 be identical to the unknown spectrum, but several options were given for each field. As the four largest peaks in the unknown and in the library spectrum are not identical in order, this illustrates that input data must not be too rigidly defined as slightly different experimental conditions may yield differing mass spectra.

Further searches were carried out with the text-search algorithms. An additional 19 "unknown" spectra were sought with both the text-search and Biemann-search algorithms from the same library. Approximately equal amounts of effort were required to construct the search strategies for each method. In general, the results were quite similar. In most cases, the text search yielded three or fewer citations, whereas the Biemann search was either the first or second closest match out of twenty possible matches. There were three cases in which the "unknown" compound was not present in the data base; in two of the three cases no citations were found by the text search, whereas the remaining case yielded six citations of compounds of similar chemical nature to the unknown. For these three cases of library-absent compounds, the Biemann algorithm predicted five possible matches in one case, and twenty possible matches in each of the other two cases. The search results are tabulated in Table 2.

TABLE 2

Comparison of text-search and Biemann-search algorithms

Unknown compound number	Number of text search citations	Number of Biemann search citations	Number of correct citation (Biemann)
1	1	20	1
2	1	20	2
3	2	20	6
4	8	20	1
5	5	5	1
6	4	20	1
7[a]	0	5	none
8	3	20	11
9	6	20	1
10	1	20	1
11	1	20	1
12	3	15	1
13	3	20	3
14[a]	0	20	none
15	5	20	1
16	5	20	1
17	3	20	1
18	2	20	1
19[a]	6	20	none

[a]Unknown compound not in library

A close examination of Fig. 4 yields an error. The Wiswesser structural information codes indicate that the compound benzyl acetate, contains: a carbonyl (V); oxygen (O); a phenyl group (R); a carboxyl group (VO); a methoxy group (O1); and a methyl group (1). Benzyl acetate does not contain a methoxy group. The file that was used to generate the library does not contain the original Wiswesser Line Notation (WLN) for any entry, but this information is partially supplied in the form of the Wiswesser structural information codes. Benzyl acetate has the WLN of 1VO1R which indicates, when compared with the codes in Fig. 4, that the codes were all derived from the WLN. The Wiswesser structural information code O1 is defined as a methoxy group, but a methoxy group is defined as being present in the WLN only if the character immediately succeeding the O1 is a blank or an ampersand. This would indicate that there was some error introduced in the generation of the Wiswesser structural information codes.

Errors such as the one found in Fig. 4 appear with significant frequency; consequently, it was decided that a detailed error analysis was necessary. In every citation there is some redundant information which can be used to verify portions of the data internally. An example of how the error analysis is accomplished may be illustrated as follows. If a V is present in the WN field (carbonyl presence), then there should be at least one oxygen (*O) present in the FO field. The number of citations with a V(WN) present and a *O(FO)

absent totals 13 in the entire data base of 18,806 entries. As can be seen from this example, the method used is to find two different sources of redundant information in the same citation and compare them. If the information agrees in context, the citation is consistent; if the information differs in context, the citation is inconsistent and must contain an error. The example presented here illustrates errors in the WLN structural information codes, but the error detection in redundant information is not limited only to this field. Any two or more fields that contain redundant information may be examined, e.g. name versus partial molecular formula. This error detection method has been applied to the ASTM Infrared Spectral Index of 91,875 entries where significant errors were elucidated [23]. The results of the error analysis on the mass spectral library are presented in Table 3.

Examination of Table 3 shows that most of the errors are infrequent and insignificant; however, there are a few which occur rather often. For example, errors occur when bromine (E) and fluorine (F) are present in the WN field, but are not present in the FO field. E and F can be substituent ring locants in WLN and correspond to the *meta* and *ortho* positions, respectively, on a phenyl ring, or as the fifth and sixth substituent positions on other ring systems. These codes may have been incorrectly extracted for bromine and fluorine. Overall, the error rate is low when checked in this manner; however, as the complete WLN was unavailable, precise examination of the WN field was not

TABLE 3

Error tabulation

DATA PRESENT	DATA ABSENT	NUMBER OF ERRORS
V (WN)	*O (FO)	13
Q (WN)	*O (FO)	9
O (WN)	*O (FO)	14
VO (WN)	*O (FO)	3
O1 (WN)	*O (FO)	3
O2 (WN)	*O (FO)	1
O3 (WN)	*O (FO)	2
Z (WN)	*N (FO)	5
M (WN)	*N (FO)	135
N (WN)	*N (FO)	41
WN (WN)	*N (FO)	0
K (WN)	*N (FO)	89
S (WN)	*S (FO)	43
WS (WN)	*S (FO)	3
E (WN)	*E (FO)	693
*E (FO)	E (WN)	7
F (WN)	*F (FO)	482
*F (FO)	F (WN)	4
G (WN)	*G (FO)	94
*G (FO)	G (WN)	19
*S (FO)	S (WN)	23

possible. This illustration is not intended to demonstrate the high error rate often attributed to hand-generated WLN codes; instead, it is intended to demonstrate the ability of the text-search algorithms to check errors. Undoubtedly, special programs may be written in high level languages to find errors in data bases, but with the general text-search algorithms, the error-checking feature is an inherent capability.

Search times for the system cannot be determined accurately as the search depends on the complexity of the search profile. A single search on the large library file (all 27 fields) requires a minimum of about 12 min; the same search on the short library requires only about 7 min. Batch processing improves the search time per profile and hence the performance. For example, 53 complex profiles were run on the shorter library file in 43 min 24 s, i.e. ca. 50 s per profile, which is a respectable search time. This search time is very similar to that for the Biemann search algorithms. As stated above, the same set of data (19 "unknown" spectra) were run on both the text search and Biemann algorithms. The actual CPU time for the Biemann search was 75 s on an IBM Model 370/System 155 computer; however, the turn-around time (job-submission to output pick-up) was nearly identical with the text search run. This will generally be the case when searches are run on a system which is not interfaced and dedicated to a mass spectrometer.

CONCLUSIONS

The minicomputer search system described is extremely versatile, which more than compensates for its slowness. A system is currently being developed to alleviate the problem of slowness. An Interdata 6/16 central processing unit with a 600-ns cycle time and a 65-Kbyte MOS memory has been purchased specifically for text searching. The CPU also has as peripherals two nine-track tape drives (800 bpi 45 ips and 1600 bpi 125 ips). With the computer architecture of the Interdata and the faster tape drives, search times should be cut by a factor of 5—10. The text search is flexible and can be tailored to the needs of individual users by setting the profile input criteria as narrow or as broad as required. Unlike some other mass spectral search systems, additional information not directly derived from the spectra may be incorporated. The search system also makes it possible to cross-check the data internally. In this way, erroneous citations may be flagged and corrected; other fields may be shown to contain unreliable data and hence be undesirable for use in search strategies. Furthermore, the search algorithms are general and may be used on virtually any data base, once that data base has been converted to the appropriate format. In the time that this mass spectral search has been operational, hundreds of unknown spectra have been efficiently and effectively searched. Through the overall versatility of the system, experienced users have proved it can consistently produce concise, accurate results.

The authors are greatly indebted to Professor F. W. McLafferty, Cornell University, for supplying a copy of his mass spectral data file, and thank Mr. G. E. Marshall for his help in compiling search system statistics. The financial support of the National Science Foundation is gratefully acknowledged. This paper was presented in part at the 27th Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, Cleveland, 1976.

REFERENCES

1 S. Abrahamsson, S. Stenhagen-Stallberg and E. Stenhagen, Biochem. J., 92 (1964) 2 p.
2 S. Abrahamsson, Sci. Tools, 14 (1967) 29.
3 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1464.
4 R. A. Hites and K. Biemann, Adv. Mass Spectrom., 4 (1968) 37.
5 B. A. Knock, I. C. Smith, D. E. Wright, R. G. Ridley and W. Kelly, Anal. Chem., 42 (1970) 1516.
6 H. S. Hertz, R. A. Hites and K. Biemann, Anal. Chem., 43 (1971) 681.
7 P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 690.
8 S. L. Grotch, Anal. Chem., 42 (1970) 1214; 43 (1971) 1362; 45 (1973) 2; 46 (1974) 526.
9 L. E. Wangen, W. S. Woodward and T. L. Isenhour, Anal. Chem., 43 (1971) 1605.
10 P. R. Nageli and J. T. Clerc, Anal. Chem., 46 (1974) 739A.
11 F. Erni, J. T. Clerc and S. Hishida, Kagaku No Ryoiki, 71 (1974) 110; Chem. Abstr., 81: 70003z (1974).
12 T. Ö. Grönneberg, N. A. B. Gray and G. Eglinton, Anal. Chem., 47 (1975) 415.
13 K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty, J. Am. Chem. Soc., 95 (1973) 4185.
14 N. A. B. Gray and T. Ö. Grönneberg, Anal. Chem., 47 (1975) 419.
15 S. R. Heller, H. M. Fales and G. W. A. Milne, Org. Mass Spectrom., 7 (1972) 107.
16 S. R. Heller, Anal. Chem., 44 (1972) 1951.
17 S. R. Heller, D. A. Koniver, H. M. Fales and G. W. A. Milne, Anal. Chem., 46 (1974) 947.
18 S. R. Heller, R. J. Feldman, H. M. Fales and G. W. A. Milne, J. Chem. Doc., 13 (1973) 130.
19 S. L. Grotch, Anal. Chem., 47 (1975) 1285.
20 G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer and F. W. McLafferty, Anal. Chem., 48 (1976) 1362.
21 T. L. Isenhour, W. S. Woodward and S. R. Lowry, J. Chem. Inf. Comput. Sci., 15 (1975) 115.
22 H. B. Woodruff, S. R. Lowry and T. L. Isenhour, J. Chem. Inf. Comput. Sci., 15 (1975) 207.
23 J. A. de Haseth, H. B. Woodruff and T. L. Isenhour, Appl. Spectrosc., 31 (1977) 18.

Analytica Chimica Acta, 103 (1978) 121—132
Computer Techniques and Optimization

AN APPROACH TO AUTOMATED PARTIAL STRUCTURE EXPANSION†

C. A. SHELLEY, T. R. HAYS and M. E. MUNK*
Department of Chemistry, Arizona State University, Tempe, Arizona 85281 (U.S.A.)
R. V. ROMAN
Department of Mathematics, Arizona State University, Tempe, Arizona 85281 (U.S.A.)
(Received 18th December 1977)

SUMMARY

An algorithm (ASSEMBLE) to construct all structures consistent with the structural implications of the chemical and spectroscopic properties of an unknown molecule is described. The design of ASSEMBLE takes cognizance of the need to supply some non-overlapping substructure information in addition to the molecular formula, and the use of structural constraints that cannot be directly expressed as non-overlapping fragments. ASSEMBLE employs several heuristics (rules) intended to avoid the assembly of identical (isomorphic) graphs. To provide a non-redundant list of structures, duplicate structures are recognized and removed by a naming algorithm. ASSEMBLE also perceives different π-resonance forms as identical structures even when they are topologically non-equivalent.

Two important approaches to the structure elucidation of organic compounds derived from natural sources (biomolecules) and synthetic transformations are: (1) single crystal x-ray analysis and (2) the reduction of the chemical and spectroscopic properties of an unknown to its molecular structure. The latter approach combined with computer assistance to produce all structures consistent with the evidence will considerably improve the manual process.

Certain common features of the science and art of the manual process of structure elucidation may be deciphered.

1. The inference of structural implications from chemical and spectroscopic data.
2. The process of analyzing the combined structural implications to deduce a partial structure. The partial structure is comprised of known substructures, atoms that are unaccounted-for, and other structural information not specific to atoms or fragments. Thus, the partial structure summarizes the status of the problem at any given stage.
3. The partial structure or molecules compatible with it guides the design of new experiments.

†Presented in part at the 172nd National American Chemical Society Meeting Division of Computers in Chemistry, San Francisco, CA, August, 1976.

Iteration of the above process ultimately leads to the correct structure assignment. The overall process is illustrated in Fig. 1.

Attention is focussed here on three of the major components necessary to develop a computer model of the structure elucidation process: (1) the reduction of chemical and spectroscopic data to their structural implications [1, 2]; (2) the expansion of a partial structure to all molecular structures consistent with it; and (3) spectral simulation of molecular features [3]. The current status of CASE (Computer-Assisted Structure Elucidation) is summarized in Fig. 2. This paper is concerned with the molecule assembler (ASSEMBLE) which expands the partial structure to all molecules consistent with it and with any other information available. Molecule assembly embraces several disciplines, including graph theory, computer science and chemistry.

The problem of discerning the topological (constitutional) properties of a molecule is addressed by means of graph theory. In this context of topological examination, a chemical structure may be viewed as a graph. A graph consists of a set of nodes (atoms) and a set of edges (bonds) connecting the

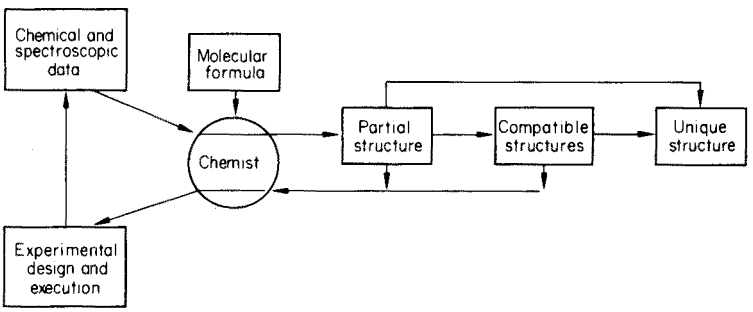


Fig. 1. "Manual" structure elucidation.

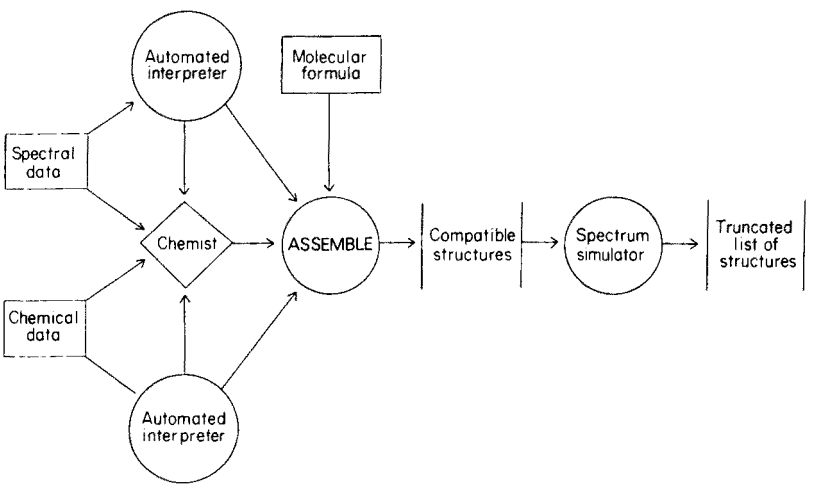


Fig. 2. CASE network.

constituent nodes. (Multiple bonds are treated as multiple edges in this paper.) An elementary cycle is a sequence of nodes, A_1, A_2, ..., A_n where there exists an edge between each pair (A_i, A_{i+1}), where $A_1 = A_n$, and where no node except the first occurs more than once, e.g., norbornane contains two cycles with five nodes and one containing six. Excellent in-depth treatments of graph theory [4] and applications in chemistry [5] are available.

Chemical graphs can be represented in computer form by a connectivity table showing explicit connections, or in a multitude of different symbolic representations, e.g., the well-known Wiswesser Line Notation (WLN). For molecule assembly, the necessity of explicit connections mandates the connection table form.

Of particular relevance in the computer science field is the area termed "artificial intelligence". Artificial intelligence programs are designed to simulate human abilities in problem solving. Applying the concept to molecule assembly involves, in part, an attempt to emulate the mental process normally used in deducing pathways to complete structures. However, in practice a technique termed heuristic search is typically used in such programs. Heuristics are "rules of thumb" or strategies which attempt to limit drastically the search for solutions to a problem. Heuristic search does not guarantee a solution as does an algorithm; rather it offers a practical device to solve most problems. Information on artificial intelligence [6] and problem solving techniques [7] in this domain is available.

At the heart of the partial structure-expansion problem is the assembly of isomeric structures from the component atoms and known non-overlapping structural fragments. Several programs have been developed to construct chemical graphs. Kudo and Sasaki [8] use a technique involving a canonical representation which guides a permutational construction procedure in program CHEMICS (Combined Handling of Elucidation Methods for Interpretable Chemical Structures). Program MASS (Mathematical Analysis and Synthesis of Structures) generates all possible connectivity combinations by a permutational approach, and then checks that completed structures are both unique and connected, i.e., only one molecule for the formula [9].

The DENDRAL procedure [10] generates all possible ring systems by means of a vertex graph (graphs containing only nodes of degree three or higher which cannot be divided into two parts by scission of one edge) library [11] and also generates all possible acyclic substructures. The combination of these fragments in all unique pathways produces an exhaustive isomer list. Although this algorithm is limited to assembling structures from component "atoms", known non-overlapping structural fragments termed "superatoms" can be utilized. Duplicate structures, which may be formed in expansion of "superatoms" to their full identities, are removed by generating a unique name.

ASSEMBLE was designed to mimic, in part, the process of structure elucidation as practiced by chemists. With this in mind, two program requirements suggest themselves. First, to serve the intended purpose of providing

a manageable list of possible structures to chemists, some substructure information (polyatomic fragments) derived from spectroscopic and/or chemical data must be provided as input. Although the present program is capable of expanding the molecular formula alone, the DENDRAL algorithm appears to be more efficient in solving that particular problem for large molecules. By narrowing the focus of ASSEMBLE to partial structure expansion, rather than molecular formula expansion, an efficient and compact algorithm has been developed. Secondly, in "real world" problems, much useful information is generally derived from chemical and spectroscopic data which cannot be directly expressed as polyatomic fragments, e.g., the required appearance of a functional group in a ring of specified size. Four other programs which expand the partial structure under constraints have been reported: MASS [9], CONGEN [12], STR-3 [13] and CHEMICS [8]. The efficient use of computer resources requires that such information be used to constrain the molecule assembler rather than to prune invalid structures retrospectively from a larger list. The design of ASSEMBLE takes cognizance of this need.

Overview of ASSEMBLE

ASSEMBLE uses a heuristic depth-first search to expand the partial structure to the complete list of compatible molecules. It systematically expands the most recently constructed partial structure (node) first. Each node is immediately evaluated for consistency with the chemical constraints. The search space traced by ASSEMBLE is represented by a tree structure. Figure 3 illustrates this depth-first tree expansion by a sequence of "snapshots". The original partial structure problem is called the "root" node. Expansion of the root node results in "descendent" nodes, i.e., simplified partial structure problems. This process, applied recursively, leads eventually to a

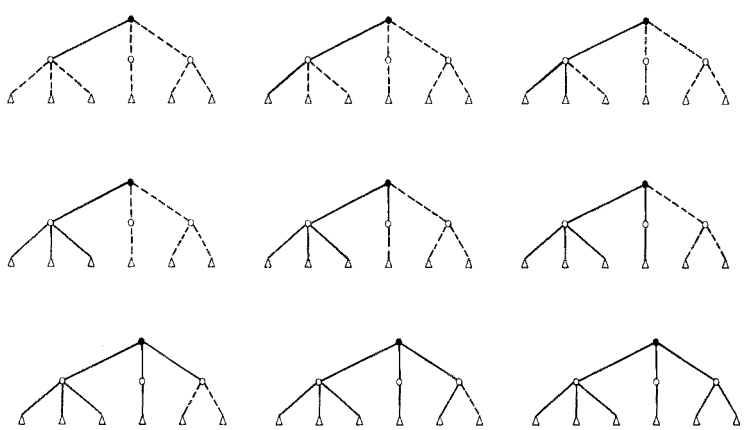


Fig. 3. Depth-first tree expansion. —— Path to expanded node. ------- Path to unexpanded node. • Root node (partial structure). ○ Descendent node (simplified partial structure). △ Terminal descendent node (molecule).

multitude of trivial partial structure problems where further expansion will form isomeric structures. The overall process is a recursive simplification or expansion of the partial structure. Thus, ASSEMBLE performs two basic steps: (1) expansion of a partial structure to descendent nodes with one bond connections, and (2) evaluation of a partial structure for compatibility with the structural constraints.

Topological symmetry

ASSEMBLE employs several heuristics which are intended to avoid the assembly of identical (isomorphic) molecular graphs. Many of these heuristics require the perception of topological (constitutional) symmetry. An algorithm which perceives topological symmetry has been described [14, 15]. The approach consists of three steps: (1) the non-hydrogen atoms are partitioned into classes by associating with each atom a list of properties; (2) tests are made to see if the atoms in a class can be partitioned further by examining the classes to which their nearest neighbors belong; (3) step (2) is repeated if it results in more classes, otherwise the process is stopped. (A similar approach, which does not claim complete differentiation, has been published [16].) For partial structures, the property list in step (1) consists of five topological properties. Four are local properties: the number of two-electron covalent bonds joining an atom to non-hydrogen atoms, elemental type, free valence and the atom-tag-constraint descriptor. (Available bonding sites are referred to as "free valences". Atom-tag constraints define atom environments without concern for overlapping structural fragments, hence this descriptor designates class membership as specified by atom tags.) In addition, a cycle list may further partition atoms for cyclic molecules. For each atom in a molecule of n atoms, the list is comprised of the number of distinct elementary cycles, of each size, in which the atom occurs, starting with those of length 3 and increasing to those of length $n - 1$. While no proof has been developed to ensure that the algorithm will operate correctly on all graphs, the algorithm has correctly partitioned the atoms in each molecule from a substantial collection of organic graphs as well as numerous contrived graphs [17] which have served as counter examples to other algorithms.

DETAILED DESCRIPTION OF THE ASSEMBLE PROGRAM

A simple partial structure comprising one substructure and two "unaccounted-for atoms" is portrayed by (I).

Molecular formula:	$C_5H_{11}N_1$	
Bonding Fragment:	$^1C—^2C=^6N—^3C$	(I)
Unaccounted-for-atoms:	$^4C\ ^5C$	

Partial structures are represented in the computer by a connection table. Atoms with free valences at the root node receive the smallest sequence numbers. Atoms with free valences in the same fragment are numbered

sequentially. An additional restriction on the assignment of sequence numbers is also necessary to avoid duplicate construction whenever the original partial structure is symmetrical. Assignment preference for atom one is given to any atom in an equivalence class with fewest members. For example, the two unaccounted-for atoms of partial structure (I) would not be assigned sequence number one. In general, many sequence number assignments are possible and equally suitable. The computer representation of partial structure (I) is shown in Table 1. Hydrogen atoms are never directly involved in bonding; rather, sufficient free valences are left unoccupied to accommodate them. In the example, 11 hydrogen atoms imply that no additional unsaturations (rings or multiple bonds) are allowed.

Partial structure expansion to complete molecules is assured by initially selecting a "bonding fragment". ASSEMBLE always expands a partial structure by elaborating on the bonding fragment. Thus, as expansion proceeds, the bonding fragment increases in complexity until it becomes a single complete molecule, i.e., an atom in the bonding fragment is always involved in bonding. In addition, the use of a bonding fragment makes the recognition of ring formation a trivial step since only connections within the bonding fragment result in ring formation. By convention, the fragment containing atom number one is designated as the bonding fragment. The bonding fragment status of each atom with free valence is contained in the ATOMSTAT (atom status) array. A "0" indicates the atom is contained in the bonding fragment and "1" specifies the opposite. A change in an atomic bonding fragment status is reflected at the descendent node by the corresponding ATOMSTAT modification. ATOMSTAT values for partial structure (I) are shown in Table 1.

To ensure that all bond combinations are attempted in the expansion of each partial structure, atoms are designated as bond "initiating" and "terminating". The initiating atom is always contained in the bonding fragment. By convention, a local pointer to the initiating atom always starts at atom number one and increases until no additional atoms with a larger sequence number and free valence exist. A local pointer also specifies the current

TABLE 1

Computer representation of partial structure (I)

Atom sequence number	Connection Table			Free valence	ATOMSTAT
	Connected to	Number of connections	Element type		
1	2	1	C	3	0
2	1, 6, 6	3	C	1	0
3	6	1	C	3	0
4	—	0	C	4	1
5	—	0	C	4	1
6	2, 2, 3	3	N	—	—

terminating atom. To avoid the construction of duplicates, the initial value of this pointer varies. It is always initialized to one at the root node. However, when the initiating atom pointer at a descendent node is equal to the initiating atom pointer of the parent node, the terminating atom pointer at the descendent node is also initialized with the present value of the terminating atom pointer of the parent node plus one.

With the specified representation of the partial structure (Table 1) and bonding restrictions, a simple combinational approach leads to the exhaustive and complete list of candidate structures. Recalling that no additional unsaturations are allowed, atom 4 could be connected to atoms 1, 2 and 3. A connection between 4 and 5 is forbidden because the fragment must always be elaborated upon. Although connections of atom 5 with 1, 2 and 3 are also possible, each of these connections lead to previously assembled partial structures, because atoms 4 and 5 are topologically equivalent. To avoid this redundancy in advance, ASSEMBLE perceives the topological symmetry of the partial structure. Topological symmetry perception of a partial structure proceeds concertedly for all component fragments. Thus, ASSEMBLE recognizes the equivalence of atoms 4 and 5, leading to three simplified partial structures. Figure 4 shows the complete tree for this problem. The order of node expansion for the same problem is illustrated in Fig. 3.

After an atom has been selected to initiate a bond to every other atom while expanding a specific node, its free valence is no longer available. The ATOMSTAT value is changed to "2" to implement this restriction (bonding atoms must possess ATOMSTAT values of 1 or 0), e.g., when ASSEMBLE returns to the root node (A), the ATOMSTAT value of atom 1 becomes 2. Furthermore, each atom topologically equivalent to any other atom, at this node or any descendent node, possessing an ATOMSTAT value of 2, is also assigned an ATOMSTAT value of 2. This is substantiated by recognizing that forbidding an atom to bond to non-hydrogen atoms

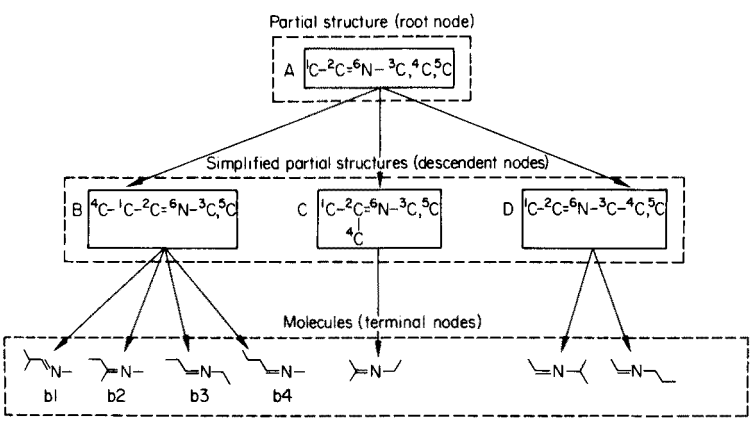


Fig. 4. Depth-first tree expansion details.

implies that the free valences are connected to hydrogen atoms. Thus, all equivalent atoms must also "bond" to hydrogen atoms.

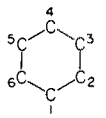
Figure 4 clarifies this rule. After the partial structure (node B) resulting from a connection between atoms 1 and 4 has been expanded, ASSEMBLE returns to expanding the root node. The terminating atom pointer is first increased to 5, but ASSEMBLE recognizes the equivalence with atom 4 and again increases this pointer. Since no additional atoms have free valence, the ATOMSTAT value of atom 1 is changed to 2, the initiating atom pointer is increased to 2, and the terminating atom pointer is again initialized at 1. ASSEMBLE recognizes that atoms 1, 2 and 3 cannot terminate a bond because no unsaturations are left and the pointer is increased to 4.

A connection between atoms 2 and 4 now results in node C and both pointers at this node are initialized at atom 1. (Because ASSEMBLE is a recursive procedure and both pointers are local variables, the pointers at node A remain unchanged.) At this time, the ATOMSTAT value of atom 4 is changed to a 0 to reflect the change in bonding fragment status from the previous connection. The subsequent perception of topological symmetry demonstrates that atoms 1 and 4 are equivalent, thus the ATOMSTAT value of atom 4 is changed to a 2 to abide by the previously stated rule. The initiating atom pointer is changed to atom 3 since atom 1 has an ATOMSTAT value of 2 and atom 2 has no free valences. After the molecule resulting from the connection between atoms 3 and 5 has been completed, the initiating atom pointer is changed to 4. Atom 4, having an ATOMSTAT value of 2, cannot initiate a bond, thus preventing the formation of structure b2 a second time, and demonstrating the effectiveness of the above heuristic. The expansion of node C is now complete and ASSEMBLE returns to expanding node A.

The expansion process increases in complexity when bond unsaturations can be formed. Even though multiple bonds are represented as multiple connections in the connection table (Table 1), such connections are formed simultaneously rather than successively. Thus, a pair of atoms, with sufficient free valences, may be connected twice (or even three times) in one step. This modification is advantageous when partial structures are expanded under constraints.

In addition, the connection process within the bonding fragment also becomes more complex when bond unsaturations are allowed. The situation can be illustrated with partial structure II and the formula C_6H_{10}. Topologically, all atoms of II are identical, i.e., choosing any site to initiate a bond is sufficient to elaborate all structures. ASSEMBLE recognizes this and allows only atom 1 to initiate a bond. Now atoms 2, 3 and 4 represent the three possible choices for bond termination since atoms 5 and 6 are equivalent to atoms 3 and 2, respectively. To detect this situation, the initiating atom is assigned to a unique equivalence class before the second step of the topological symmetry procedure is started. The classes are now

partitioned by the topological symmetry algorithm into new classes which guide the selection of the terminating atom.



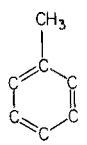
(II)

Even though many efficient heuristics are used to avoid prospectively duplicate construction, it is still possible to form duplicate molecules. To provide a non-redundant list of structures, any duplicate structure that does result is removed by an algorithm [18] which uniquely names each structure and determines retrospectively whether it was previously generated.

RECOGNITION OF RESONANCE FORMS

ASSEMBLE perceives different π-resonance forms as identical molecular structures even when they are topologically non-equivalent. Examples of topologically non-equivalent resonance forms are the two Kekulé structures of *o*-xylene. *m*-Xylene has two resonance forms which are topologically equivalent. Topologically non-equivalent resonance forms are excluded both prospectively by the topological symmetry algorithm [14, 15] and retrospectively by the structure-naming algorithm [18].

The first of these methods can be illustrated by considering the polyatomic fragment III, a methyl group and four hydrogens as a partial structure expansion



(III)

problem. Topologically, all atoms of fragment III are non-equivalent when only one Kekulé structure is considered. In such cases, however, the topological symmetry algorithm recognizes that the two *ortho* and the two *meta* positions are equivalent. Rather than generating five xylene structures, only three unique structures are constructed.

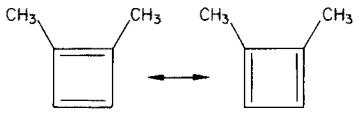
In spite of these heuristics, π-resonance forms can be constructed in some cases. For this reason, a naming algorithm was designed which produces redundant names for π-resonance forms to provide a non-redundant list of structures. The algorithm produces identical names for structures which are topologically non-equivalent, but differ only in the location of π-electrons. However, the algorithm does not distinguish between conjugated cyclic structures that do and do not obey Hückel's rule ($4n + 2\ \pi$ electrons, $n = 0, 1, 2 \ldots$). Consider the non-conforming resonance structures IV and V. Although these structures would be highly unstable, they are topologically

TABLE 2

Structural isomers for various molecular formulas
(Numbers in parentheses were obtained from the DENDRAL program [19]. The differences between ASSEMBLE and DENDRAL are due to removal of equivalent resonance structures by ASSEMBLE. See text for a discussion of resonance perception.)

Composition	Degree of unsaturation							
	0	1	2	3	4	5	6	7
C_5	3	10	26	40	38	20	6	
					(40)	(21)		
C_4O_1	7	26	55	62	34	6		
					(36)	(7)		
C_4N_1	8	35	85	115	83	26		
				(116)	(87)	(27)		
C_3O_2	11	34	52	34	7			
C_3N_2	14	62	136	153	83	14		
				(155)	(86)			
$C_3O_1N_1$	21	84	154	135	44			
				(136)	(46)			
C_2O_3	10	22	20	5				
C_2N_3	14	58	110	98	33			
				(99)	(34)			
$C_2O_2N_1$	28	84	99	40				
$C_2O_1N_2$	31	115	177	113	19			
				(114)	(20)			
C_6	5	25	77	158	212	177	76	15
				(159)	(217)	(185)	(85)	(19)
C_5O_1	14	74	205	336	311	145	20	
				(337)	(318)	(151)	(21)	
C_5N_1	17	100	313	590	672	419	100	
				(593)	(685)	(437)	(112)	
C_4O_2	28	122	263	300	159	26		
				(301)	(163)	(28)		
C_4N_2	38	218	633	1050	985	441	56	
				(1058)	(1005)	(465)	(64)	
$C_4O_1N_1$	56	299	764	1063	759	208		
				(1069)	(775)	(216)		
C_3O_3	28	102	152	98	16			
C_3N_3	45	259	681	964	695	185		
				(969)	(706)	(194)		
$C_3O_2N_1$	90	391	732	(639)	199			
				(641)	(202)			
$C_3O_1N_2$	102	527	1194	1363	691	85		
				(1371)	(703)	(88)		

possible and differ only in the location of π-electrons, consequently, they



IV V

would be perceived as identical by the naming algorithm.

MOLECULAR FORMULA EXPANSION

Although ASSEMBLE was designed to expand partial structures systematically under constraints, the molecular formulae listed in Table 2 were expanded to structural isomers to substantiate the exhaustive character of the process. These results are consistent with the published results from the DENDRAL algorithm [19]. In those cases where a larger number of structures are reported by DENDRAL (the numbers in parentheses in Table 2), it was shown manually that the excess was excluded by ASSEMBLE as equivalent to others on the basis of resonance. Thus, the molecular formula $C_4O_2H_2$ has four resonance pairs (Figure 5).

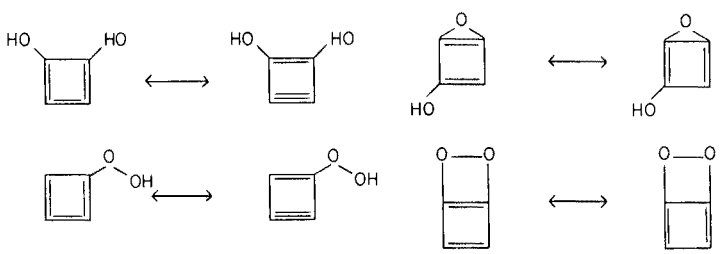


Fig. 5. Topologically non-equivalent resonance pairs for the molecular formula $C_4O_2H_2$.

The authors gratefully acknowledge the support of this project by the National Institutes of Health (GM 21703) and Arizona State University's Computer Center.

REFERENCES

1 H. B. Woodruff and M. E. Munk, J. Org. Chem., 42 (1977) 1761.
2 C. A. Shelley, et al., in D. H. Smith (Ed.), Computer-Assisted Structure Elucidation, A.C.S. Symposium Series, Vol. 54, 1977, p. 92.
3 C. A. Shelley and M. E. Munk, Anal. Chem., submitted.
4 F. Harary, Graph Theory, Addison-Wesley, Reading, MA, 1969.
5 A. T. Balaban, Chemical Applications of Graph Theory, Academic Press, New York, NY, 1976.
6 E. A. Feigenbaum and J. Feldman, Computers and Thought, McGraw-Hill, New York, NY, 1963.

7 N. Nilsson, Problem Solving Methods in Artificial Intelligence, McGraw-Hill, New York, NY, 1971.
8 Y. Kudo and S. Sasaki, J. Chem. Inf. Comp. Sci., 16 (1976) 43.
9 V. V. Serov, M. E. Elyashberg and L. A. Gribov, J. Mol. Struct., 31 (1976) 381.
10 L. M. Masinter, et al., J. Am. Chem. Soc., 96 (1974) 7702.
11 R. E. Carhart et al., J. Chem. Inf. Comp. Sci., 15 (1975) 124.
12 R. E. Carhart et al., J. Am. Chem. Soc., 97 (1975) 5755.
13 B. D. Cox, Computer Program STR-3, Ph-D. Dissertation, Arizona State University, 1973.
14 C. A. Shelley and M. E. Munk, J. Chem. Inf. Comp. Sci., 17 (1977) 110.
15 C. A. Shelley, M. E. Munk and R. V. Roman, J. Chem. Inf. Comp. Sci., submitted.
16 D. G. Corneil and C. C. Gotlieb, J. Assoc. Comp. Mach., 17 (1970) 51.
17 D. G. Corneil, Computer Science Department, University of Toronto, personal communication.
18 C. A. Shelley, M. E. Munk and R. V. Roman, in preparation.
19 D. H. Smith, J. Chem. Inf. Comp. Sci., 15 (1975) 203.

Analytica Chimica Acta, 103 (1978) 133—140
Computer Techniques and Optimization

MISSING VALUES IN TIME SERIES AND THE IMPLICATIONS ON AUTOCORRELATION ANALYSIS

C. B. G. LIMONARD

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

(Received 28th September 1977)

SUMMARY

Arising from a practical situation, the effect of missing values on autocorrelation analysis of first-order autoregressive stochastic stationary processes is investigated. A practical solution is described, and Monte Carlo simulations are performed to test the validity and applicability of the procedure. The procedure is seen to be valid and applicable for other than extreme situations.

A problem rarely dealt with in statistical literature is that of missing values in time series analysis and in particular its implication on autocorrelation analysis. Often, however, data cannot be collected in an ideal way. Part of an investigation concerning dynamic aspects of quality control systems in clinical chemistry laboratories in The Netherlands [1] involved this problem. For this investigation thirty clinical chemistry laboratories participating in the "Stichting Kwaliteitsbewaking Klinisch Chemische Ziekenhuislaboratoria" submitted their intralaboratory serum calcium and urea quality control data over the period 1974/1975. These laboratories not only use a wide range of different analytical techniques for these two determinations but also use a wide variety of quality-control systems. The working hypothesis was that these laboratories could be represented as first-order autoregressive stochastic stationary processes, and so the current investigation was restricted to such processes.

A possibility for identifying processes from discrete time series is offered by the autocorrelation function. This paper describes a procedure for the identification of first-order autoregressive stochastic stationary processes from discrete time series with missing values, when the number of missing values is taken into consideration.

THEORY

Throughout this paper only first-order autoregressive stochastic stationary processes, henceforth called the process, are considered, with stochastic indicating that the process value x_t at time t is not fixed by a mathematical model, but is defined merely in terms of probability distributions $p(x_t)$. For

gaussian processes, $p(x_t)$ is determined by the mean value $E(x_t)$ and the standard deviation σ_{x_t} as given by Box and Jenkins [2]. When $E(x_t)$ and σ_{x_t} are constant in time, the process has a stationary behaviour. This implies that the process values, represented as a discrete time series, are merely functions of the intervals between the successive observations.

From limited time series of N successive measurements, $E(x_t)$ and σ_{x_t} are estimated according to

$$\bar{X} = \frac{1}{N} \sum_{t=1}^{N} x_t \quad \text{and} \quad s_{x_t}^2 = \frac{1}{N-1} \sum_{t=1}^{N} (X_t - \bar{X})^2$$

The correlations between successive observations are calculated by computing the autocovariance estimates c_k

$$c_k = \frac{1}{N-k-1} \sum_{t=1}^{N-k} (X_t - \bar{X}) \cdot (X_{t+k} - \bar{X}), \quad k = 0, 1 \cdots, M \tag{1}$$

where k is the time lag expressed in units of sampling interval and N the length of the series in the same units. Autocorrelation estimates r_k are then obtained from

$$r_k = c_k / c_0, \quad k = 0,1 \cdots, M \tag{2}$$

For the processes investigated, the autocorrelation function is a continuously decreasing function described by $r_k = \exp(-k/T_x)$ for $k \geqslant 0$, in which the time constant of the process, T_x, is a measure of the frequencies occurring in the process. The variance in the autocorrelation estimates calculated from N observations is approximated by eqn. (3), derived by Bartlett [3]

$$\text{var}(r_k) = \frac{1}{N} \left\{ \frac{(1+\rho^2) \cdot (1-\rho^{2k})}{(1-\rho^2)} - 2k\rho^{2k} \right\} \text{ for } k \ll N \tag{3}$$

with $\rho = \exp(-1/T_x)$.

Problem

Consider the situation in which a process is sampled at regular time intervals Δ_t. The discrete time series $Z_1, Z_2, \cdots, Z_N$ of N successive observations is then regarded as a sample realization from an infinite population of such series, representative of the process.

Now suppose that, for some reason, not all N measurements are available but merely N^1, during the same total realization period. Denoting the structure of the known and unknown terms by Z and ?, there could raise, for example the following sequence of observations

ideal situation : Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
with missing values: Z_1 ? ? Z_4 ? Z_6 Z_7 Z_8 ? ?

In the case of a single gap of one or more observations, with long unbroken sequences immediately before and after the gap, interpolation is possible [4]. Several gaps, provided they are well separated, can also be filled or interpolated

separately. Calculating the autocorrelation function need not be a problem subsequently.

In the situation with a lot of missing data and no long unbroken sequences before and after the gap, as considered above, interpolation is difficult, if possible at all.

Solution

Let a process of realization length R be sampled at equidistant time intervals Δ_t. When the sequence is undisturbed, N successive measurements form the discrete time series $Z_1, Z_2, \cdots, Z_N$. Now assume that the series is filled with gaps and consists of N^1 observations during the same realization period R. To compute the autocorrelation function of this series, fill all gaps with an arbitrary, recognizable value (AV) not already present in the series. This guarantees the equidistance of all observations, a necessity for autocorrelation analysis. Calculate the average process value from the observations, having omitted all arbitrary values by

$$\overline{Z} = \frac{1}{N^1} \sum_{t=1}^{N} Z_t \quad \text{with } Z_t \neq AV$$

and reduce all N^1 observations by $\overline{Z}$, as a result of which the series $W_1, W_2, \cdots W_N$ is formed. Suppose the sequence (A) to be

(A) $W_1 W_2 AV AV W_5 AV W_7 AV AV W_{10} W_{11} AV W_{13} W_{14} W_{15} W_{16} AV W_{18} W_{19} AV$

Autocovariance estimates are then computed from

$$c_k = \left[\sum_{t=1}^{N-k} W_t \cdot W_{t+k} \right] \Big/ [Q(k) - 1] \quad (k = 0, 1, \ldots, M) \tag{4}$$

in which multiplications involving an arbitrary value (AV) are omitted (i.e. $W_t \neq AV$ and $W_{t+k} \neq AV$), and where $Q(k)$ denotes the number of multiplications in which no arbitrary value is involved, for the various time lags k.

The usual procedure would be to replace the missing values by the overall mean $\overline{z}$. In the series, however, some observations may actually equal $\overline{z}$. Replacing a missing value by $\overline{z}$ thus makes a true observation of the series indistinguishable from a missing value. In the first case, $Q(k)$ would remain unchanged, but in the latter situation $Q(k)$ should be decreased by 1 each time, in the multiplication of eqn. (4), where an arbitrary value is involved. Therefore, for strictly computational reasons, all missing values are replaced by a value which can be chosen freely, that is recognizable by the computer. For the example (A)

k	0	1	2	3
$Q(k)$	12	6	5	7

With no gaps in the sequence $Q(k)$ running from 20—17. In general, $Q(k) = N - k - L\ (k)$ where N is the realization length/Δt, k the time lag, and $L(k)$ the number of multiplications involving an arbitrary value. The

autocorrelation estimates r_k are then computed from eqn. (2). From eqn. (3) it follows that in a sequence with no missing values, var(r_k) where $k = 1, 2, \cdots, M$, is always computed with constant number of observations N if $k \ll N$. With missing values in the time series, the number of observations varies with each autocorrelation estimate. The variance in these estimates were calculated by replacing N in eqn. (3) by $N(k)$

$$\text{var}(r_k) = \frac{1}{N(k)} \left\{ \frac{(1+\rho^2)\cdot(1-\rho^{2k})}{(1-\rho^2)} - 2\,k\rho^{2k} \right\} \tag{5}$$

with $N(k)$ denoting the number of observations actually involved when computing r_k, $k = 0, 1, \cdots, M$ according to eqn. (2) following eqn. (4). For the example (A), leading to

k	0	1	2	3
$N(k)$	12	10	8	11

Simulations were performed to test the validity of the procedure.

Simulations

Computations and statistics. First-order stochastic stationary processes were simulated with a discrete "white-noise" generator from the IBM library. With x representing a discrete white-noise process with zero mean, variance σ_x^2 equal to 1, and $E(x_n \cdot x_m) = 0$ for $m \neq n$, a Markov process is generated according to

$$Z_{n+1} = a \cdot Z_n + b \cdot x_{n+1} \tag{6}$$

where $a = \exp(-1/T_x)$ and $\sigma_Z = b \cdot (1-a^2)^{-1/2}$ following Naylor et al. [5] and Gelb and Palosky [6].

The series $Z_1, Z_2, \cdots, Z_N$ of N successive terms according to eqn. (6) were then compared with gaps containing sequences of the same realization length ($N.\Delta t_{\text{sampling}}$) $y_1, y_2, \cdots, y_N$, submitted by the participants taking part in the investigation mentioned. A term of the Z-series was thus omitted and replaced by an arbitrary recognizable value when the observation at the sample place in the y-series was not present. If not, the Z-value remained unchanged. Thus a series $Z_1, \cdots, Z_N$ was formed containing N^1 values and $N - N^1$ missing values.

According to the procedure described, autocorrelation estimates and the variance in these estimates were calculated for $k = 1, \cdots, 20$, for both original "gap-free" series Z as well as "gap-containing" series Z. A weighted least-squares fitting procedure, based on an iteration method developed by Meiron [7] was used to estimate the fitting function $r_k = \exp(-k/T_x)$, $k = 0, \cdots, 20$, for both performances. The curve-fitting program itself was based on the iteration formula $T_{m+1} = T_m - (B_m + pC_m)^{-1} \cdot G_m$, where T_{m+1} is the value of parameter T of the model, calculated in the $(m+1)$th iteration; B_m is the matrix of the partial derivatives of the fitting function using the parameter value from the mth iteration, C_m as matrix B_m, but with

the off-diagonal elements zero; p is the damping constant and G_m the matrix of the residual differences between observed and calculated data points from the mth iteration.

The reduced chi-square statistic X_ν^2, defined as the ratio of the estimated variance of the fit s^2 to the parent variance σ^2 (times the number of degrees of freedom) and described by Bevington [8], was used as the goodness of fit criterion. The standard deviation σ_{T_x} of the fit parameter T_x was calculated by the error propagation expression described in many textbooks [e.g. 8].

RESULTS

Table 1 shows the effect of missing values in time series analysis on $Q(k)$ and $N(k)$ when r_k and $\mathrm{var}(r_k)$ $k = 0, 1, \cdots, 20$ are calculated for the performed simulations. These are 2 extreme situations. With the sequence $N = 250$ and $N^1 = 99$, a series arose in which many observations at the beginning were followed by a gap followed by very few missing values towards the end of the series.

The sequence with $N = 1992$ and $N^1 = 218$ was the result of the following practical situation. A process was sampled during a maximal 3 h per day at a sampling rate of 1 h during 83 consecutive days. Calculating autocorrelation estimates calls for equispaced observations. The procedure now described warrants this. Suppose the process had been sampled during 24 h, resulting in 24 observations. With only 3 observations on one day, and, say, also 3 for the next, proceed as follows. Insert $24 - 3 = 21$ arbitrary recognizable values not already present in the series, leading to the sequence $Z_1\ Z_2\ Z_3\ (AV)_{21}\ Z_{25}\ Z_{26}\ Z_{27}\ (AV)_{21} \cdots$. Autocorrelation estimates with $2 < k < 21$ cannot be calculated from eqn. (2) following eqn. (4). The autocorrelation function is thus to be calculated from only 3 data points, $k \ll 2$. $N(k)$ and

TABLE 1

Effect of missing values on $Q(k)$ and $N(k)$ in simulated time series with N^1 observations and $N - N^1$ missing values

N	130		250		424		587		1992	
N^1	87		99		338		400		218	
k	$Q(k)$	$N(k)$	$Q(k)$	$N(k)$	$Q(k)$	$N(k)$	$Q(k)$	$N(k)$	$Q(k)$	$N(k)$
0	87	87	99	99	338	338	400	400	218	218
1	66	86	18	26	269	334	303	390	77	148
2	48	82	51	81	257	324	230	381	10	20
3	47	64	36	62	261	321	235	318	—	—
4	47	65	35	60	263	325	232	315	—	—
5	48	81	48	77	257	328	231	382	—	—
10	45	63	35	63	256	322	231	315	—	—
15	59	81	14	26	257	332	295	388	—	—
20	57	80	13	21	253	326	292	386	—	—

TABLE 2
Estimated time constants for series with (*b*) and without (*a*) missing values

$N = 130$ $N^1 = 87$

TW = 1				TW = 2				TW = 6			
a		*b*		*a*		*b*		*a*		*b*	
T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}
1.22	0.17	1.34	0.24	1.76	0.20	1.48	0.23	4.23	0.24	4.44	0.45
1.51	0.18	0.98	0.24	1.60	0.23	2.12	0.34	6.22	0.42	7.21	0.76
0.91	0.14	0.79	0.26	1.51	0.25	1.67	0.34	3.99	0.38	3.58	0.42
0.90	0.12	1.06	0.24	1.57	0.21	2.08	0.38	8.23	0.19	7.79	0.47
0.61	0.11	0.81	0.17	1.93	0.34	2.21	0.32	4.58	0.20	4.39	0.31
0.89	0.17	0.77	0.20	1.66	0.36	1.47	0.24	4.65	0.41	4.85	0.39
0.87	0.14	1.11	0.22	2.21	0.50	2.26	0.60	4.84	1.21	5.13	1.26
0.96	0.20	0.93	0.24	1.97	0.11	1.78	0.21	6.04	0.27	6.29	0.34
0.90	0.17	1.31	0.22	1.66	0.26	1.67	0.38	4.42	0.50	4.38	0.63
1.12	0.26	1.17	0.23	1.59	0.25	1.35	0.26	7.31	0.66	7.23	0.77
$\overline{T}_x$	0.99		1.03		1.75		1.81		5.45		5.53
$\sigma_{\overline{T}_x}$	0.24		0.21		0.22		0.33		1.44		1.48

$N = 250$ $N^1 = 99$

TW = 1				TW = 2				TW = 6			
a		*b*		*a*		*b*		*a*		*b*	
T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}	T_x	σ_{T_x}
1.23	0.13	1.45	0.41	1.67	0.19	1.98	0.41	7.10	0.18	4.98	0.92
1.15	0.11	1.23	0.42	1.88	0.27	2.24	0.41	5.50	0.15	6.28	1.09
1.02	0.11	0.72	0.40	1.84	0.21	1.64	0.44	8.72	0.18	27.43[a]	19.26
0.88	0.10	0.82	0.28	2.06	0.22	2.42	0.65	5.25	0.22	3.59	0.51
0.64	0.09	1.16	0.31	2.29	0.12	1.77	0.35	8.65	0.55	12.13[a]	3.04
0.81	0.13	0.63	0.52	1.89	0.23	1.85	0.33	5.04	0.43	4.01	0.64
0.98	0.16	1.04	0.38	1.61	0.19	1.35	0.33	4.56	0.55	7.77	1.89
1.05	0.13	0.82	0.40	1.68	0.12	1.24	0.39	3.99	0.30	5.38	1.30
1.17	0.15	0.97	0.35	1.60	0.17	1.97	0.42	7.56	0.50	13.39[a]	5.70
1.07	0.10	0.76	0.33	2.33	0.25	7.17[a]	1.49	7.70	0.18	4.78	0.47
$\overline{T}_x$	1.00		0.96		1.89		1.83		6.41		5.26
$\sigma_{\overline{T}_x}$	0.18		0.26		0.27		0.38		1.74		1.42

$N = 424$ $N^1 = 337$

1.03	0.08	1.18	0.13	1.72	0.12	1.79	0.14	6.21	0.34	6.11	0.52
0.98	0.08	0.99	0.10	1.77	0.16	1.81	0.16	5.66	0.36	6.39	0.50
1.03	0.08	1.04	0.13	1.75	0.19	1.75	0.20	3.84	0.25	3.49	0.31
0.95	0.09	0.80	0.12	2.06	0.08	1.80	0.10	5.56	0.11	5.76	0.20
0.71	0.07	0.70	0.10	1.73	0.10	1.62	0.09	5.15	0.14	5.25	0.16
0.90	0.09	1.00	0.12	2.16	0.10	2.07	0.16	7.10	0.18	6.73	0.25
0.92	0.10	0.86	0.10	1.66	0.10	1.80	0.12	5.50	0.15	5.72	0.29
1.04	0.14	1.00	0.18	1.90	0.13	1.98	0.18	8.72	0.18	8.27	0.27
1.13	0.06	0.92	0.08	2.02	0.17	1.92	0.18	5.24	0.22	4.47	0.17
0.90	0.08	0.83	0.09	1.84	0.20	1.92	0.23	8.65	0.55	8.75	0.61
$\overline{T}_x$	0.96		0.93		1.86		1.85		6.16		6.09
$\sigma_{\overline{T}_x}$	0.12		0.14		0.17		0.13		1.56		1.59

$N = 588$ $N^1 = 398$

0.91	0.10	0.83	0.14	1.62	0.14	1.48	0.16	5.65	0.18	5.42	0.21
0.93	0.09	1.10	0.10	1.79	0.12	1.98	0.14	6.39	0.25	6.85	0.34
1.04	0.09	0.99	0.10	1.85	0.12	1.80	0.11	5.95	0.15	5.44	0.23
1.00	0.06	1.18	0.09	2.10	0.08	2.23	0.10	4.68	0.12	4.56	0.26
0.92	0.08	1.05	0.10	1.60	0.10	1.82	0.12	5.58	0.13	5.71	0.20
0.90	0.07	0.80	0.07	1.73	0.09	1.58	0.08	4.49	0.17	4.51	0.21
0.93	0.06	0.92	0.10	1.85	0.08	1.86	0.13	4.89	0.23	5.15	0.22
1.06	0.08	0.93	0.07	1.99	0.09	1.86	0.11	4.79	0.15	4.72	0.12
1.00	0.07	1.05	0.12	1.79	0.10	1.90	0.19	6.18	0.18	6.59	0.23
0.99	0.06	1.19	0.09	1.93	0.06	2.12	0.13	4.17	0.14	4.68	0.19
$\overline{T}_x$	0.97		1.00		1.83		1.86		5.28		5.36
$\sigma_{\overline{T}_x}$	0.06		0.13		0.16		0.22		0.77		0.83

$N = 1992$ $N^1 = 218$

1.09	0.03	1.39[a]	0.06	2.18	0.04	3.22[a]	0.23
0.98	0.04	0.96	0.22	1.85	0.06	2.10	0.50
1.00	0.04	0.70[a]	0.03	1.96	0.07	1.21[b]	0.02
1.07	0.04	1.04	0.11	1.98	0.06	2.05	0.23
1.06	0.04	1.05	0.04	2.02	0.04	2.38[b]	0.12
$\overline{T}_x$	1.04		1.02		2.00		2.18

[a] X_p^2 indicates a poorly fitting function (Probability < 0.01). These values were omitted when $\overline{T}_x$ was calculated. [b] Probability < 0.025. These values were omitted when $\overline{T}_x$ was calculated.

$Q(k)$ in this situation are therefore only given for $k \leqslant 2$ (Table 1). With the time constant used for generating the process in eqn. (6), denoted as TW, Table 2 shows the estimated time constants T_x and σ_{T_x}, according to the fit procedure described, for both series with (b) and without (a) missing values. The reduced chi-square statistic, X_ν^2, indicated a reasonable fit unless stated otherwise. From Table 2, the following observations were made.

1. The time constants calculated from time series with and without missing values are in close agreement, as summarized in Table 3. Here the discrepancy Δ between the average time constant from the series with and without missing values, i.e. $\overline{T}_x(b)$ and $\overline{T}_x(a)$, is calculated from

$$\Delta = 100[\overline{T}_x(b) - \overline{T}_x(a)] / \overline{T}_x(a) \tag{7}$$

2. At constant realization length, σ_{T_x} increases in both situations with increasing T_x, σ_{T_x} being larger when missing values are present, according to expectation.

3. With increasing number of observations in the series, with and without missing values, σ_{T_x} decreases, as expected. Furthermore, the results of the 2 extreme situations considered ($N = 130, N^1 = 87$ and $N = 1992, N^1 = 218$; Table 2) indicate possible restrictions to the procedure. In both situations the reduced chi-square test criterion indicated a poorly fitting function with X_ν^2 values with a probability of less than 0.01.

The strong decline in $Q(k)$ and $N(k)$ at $k = 1$, compared with $k = 0$ (Table 1), as a result of which r_k and $\mathrm{var}(r_k)$ are less reliable, is considered to be the main cause, valid for both situations.

The results of Table 2, for $N = 1992, N^1 = 218$, show the effect of insufficient autocorrelation estimates when calculating the autocorrelation function. In this particular situation only estimates for $k \leqslant 2$ were available. This results very quickly in significant X_ν^2-values [7].

TABLE 3

Discrepancy, Δ, between the average time constant from the series with and without missing values i.e. $\overline{T}_x(b)$ and $\overline{T}_x(a)$, calculated from eqn. (7)

N	N^1	$\overline{T}_x(a)$	$\overline{T}_x(b)$	Δ (%)	N	N^1	$\overline{T}_x(a)$	$\overline{T}_x(b)$	Δ (%)
130	87	0.99	1.03	4.04	588	398	0.97	1.00	3.09
		1.75	1.81	3.43			1.83	1.86	1.64
		5.45	5.53	1.47			5.28	5.36	1.52
250	99	1.00	0.96	4.00	1992	218	1.04	1.02	1.92
		1.89	1.83	3.18			2.00	2.18	9.00
		6.41	5.26	17.94					
424	337	0.96	0.93	3.13					
		1.86	1.85	0.54					
		6.16	6.09	1.14					

CONCLUSION

Based on the theory of first-order autoregressive stochastic stationary processes, a procedure is described which enables a direct autocorrelation analysis of discrete time series with missing values. The average time constants calculated for both series with and without missing values are in close agreement, and differ by less than 4%. In extreme situations, the procedure sometimes yields less reliable results.

The author thanks Prof. drs. G. Kateman for helpful discussion, Drs. P. Müskens for permission to use his simulation program, and T. Swenker for his contribution to this study.

REFERENCES

1 C. B. G. Limonard, J. Clin. Chem. Clin. Biochem., 15 (1977) 172.
2 G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, 1970, p. 26.
3 M. S. Bartlett, J.R. Statist. Soc., 8B (1948) 27.
4 P. Bloomfield, Fourier Analysis of Time Series, Wiley, New York, 1975, p. 243—246.
5 T. H. Naylor, J. L. Balintfy and D. S. Burdick, Computer Simulation Techniques, Wiley, New York, 1966, p. 43.
6 A. Gelb and P. Palosky, IEEE Trans. Aut. Contr., (1966) 148.
7 J. Meiron, J. Opt. Soc. Am., 55 (1969) 1105.
8 P. R. Bevington, Data reduction and error analysis for the physical sciences, McGraw-Hill, New York, 1969, p. 187.

Analytica Chimica Acta, 103 (1978) 141—149
Computer Techniques and Optimization

A SUBSTRUCTURE-ORIENTED ^{13}C-NMR CHEMICAL SHIFT RETRIEVAL SYSTEM

J. ZUPAN[§] and S. R. HELLER*
Environmental Protection Agency, Washington, D.C., 20460 (U.S.A.)
G. W. A. MILNE
National Institutes of Health, Bethesda, Md., 20014 (U.S.A.)
J. A. MILLER
Fein-Marquart Associates Inc., Towson, Md., 21212 (U.S.A.)
(Received 16th January 1978)

SUMMARY

A computer program that uses on-line generated substructures of organic compounds as input and retrieves the corresponding distributions of ^{13}C-n.m.r. chemical shifts is described and discussed. The procedure of creating the substructures and the main features of the retrieval philosophy are outlined. One search is worked out in detail to demonstrate the ability of the system.

The assigning of ^{13}C-n.m.r. spectra, i.e. the identification of the chemical shifts with the appropriate carbon atoms in the chemical environment of the molecule, is a primary application of ^{13}C-n.m.r. spectroscopy. Usually this is done by empirical rules describing the influence of neighbors on the central atom in the fragment in question, and by inspecting the spectrum to see if it fits the proposed correlation. Several tables of shift assignments have been published [1—3] since ^{13}C-n.m.r. spectroscopy became recognized as a powerful tool for the elucidation of structures of organic compounds. Normally, such tables give only the upper and lower limit of the range in which the atom under consideration is expected to give a chemical shift. Such a description is far from complete because it provides no information as to whether the distribution of chemical shifts is uniform or centered in the given interval as a normal distribution.

It transpires that the distribution of the shifts in such an interval is not normal and is very dependent on the nature of neighbors more distant than the first and second neighbors, which are usually the only ones that are considered in the manual assembly of tables of this sort.

In this paper is described a complex computer program which provides an easier and deeper insight into the distribution of chemical shifts sampled from a fairly large data collection for any specific structural environment.

[§]On leave from the Boris Kidric Chemical Institute, Ljubljana, Yugoslavia.

DATA BASE AND IMPLEMENTATION

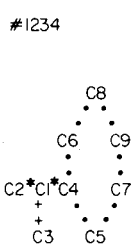
The NIH-EPA-NIC ^{13}C-n.m.r. collection [4] was used as the data base in this work. In addition to the ^{13}C-n.m.r. data, this data base contains the Chemical Abstract Registry (CAS) number, the chemical name and molecular formula of the compound. There is also associated with each compound a picture of its structure in which the atoms have been numbered. A typical display of an entry from this data base, obtained with the standard retrieval system [5, 6], is shown in Fig. 1.

The assignments of the chemical shifts were entered manually, using the same numbering as in these structures, with a specially written on-line program. Currently, the data base contains 4,024 spectra, some 2,500 of which have been assigned. Further work with the program described here to add the missing assignments in the data base is in progress.

In order to permit the user to define a chemical fragment and conduct a substructure search for it, an additional file is necessary. This file contains the connection tables of all the compounds in the ^{13}C-n.m.r. data base. The building of substructures (fragments) can be done by using the substructure program developed by Feldmann et al. [7]. The commands most frequently used in structure generation are listed in Table 1.

From the point of view of the assignment of ^{13}C-n.m.r. spectra, a very important option within the Substructure Search System is the TERMA command which allows substituents to be defined precisely. The command TERMA 3,1 for example, sets to one the numbers of neighbors of the atom with the number "3" in the query structure. It is obvious that the atom "3", whatever it is, must therefore be the last one in the chain. Without the use of the command "TERMA 3,1", structures containing atom "3" bound to 2, 3 or even more atoms would also be retrieved. Thus for the accurate definition of larger structures, the TERMA command should be used for each atom. There are, of course, many other commands for the substructure generation [8] but these are less frequently used in the present application.

The link between the Substructure Search System and ^{13}C-n.m.r. files was provided by a fast double-hashing algorithm with twin prime numbers NP and NP-2 [9] using the CAS Registry number (NUMRG) as input to obtain



Benzene, (1-methylethenyl)-
REGN= 98839 MW = 118.07
C9 H10
D.DALRYMPLE, U.OF DELAWARE, 1976.
Solvent: NEAT

SHIFT	MULT	INTENS	ASSIGN
143.5	S	0	1
141.5	S	0	4
128.4	S	0	7
127.5	S	0	9
125.7	D	0	5
112.4	T	0	3
21.8	D	0	2

Fig. 1. A typical display of the full information for the requested ID number.

TABLE 1

Most commonly used commands for generation of substructures involved in the assignment of chemical shifts. It should be noted that other commands, dealing with rings, such as ARING p1, p2 or RING n, although very valuable in other applications, are very seldom used in this case

Command	Task and description of the parameters
CHAIN n	Create a chain of n atoms.
ABRAN m AT n	Add chain (branch) of m atoms at the atom n.
SBOND n, m	Set the desired bond type between the atoms n and m. After this command is issued, the computer asks for the type of bond to be entered.
SATOM n	Set the atom type desired instead of previous choice on the position n.
TERMA n, i	Set the limit on the number of neighbors of the atom n, i in this case.
DATOM n	Delete atom n from the query structure.

the proper address, KEY, of the connection table or chemical shifts:

```
      NP = 4723
      KEY = MOD (NUMRG, NP) + 1
      INC = MOD (NUMRG, NP-2) + 2
1     CONTINUE
      KEY = KEY - INC
      IF (KEY.LE.0) KEY = KEY + NP
      .
      .
      .
      if the requested item has not
      been found on the address KEY     GO TO 1
2     CONTINUE
```

There are 21 twin prime numbers between 4000 and 5000 and the choice of 4723 was made because it is expected that about 500 new spectra will soon be added to the current data base of 4,024 spectra. About 5% free space in the address table will still be available and so unsuccessful searches will be terminated relatively rapidly.

The flow chart of the full process is given in Fig. 2. The complete search is done in three steps. First, the substructure fragment has to be built up on-line, using commands such as those shown in Table 1. Secondly, the search through the connection tables is performed in order to obtain all compounds containing the described substructure together with the numbering of the atoms in the structure as it appears in the ^{13}C-n.m.r. file. Thirdly, the chemical shifts of the retrieved compounds are inspected and those produced by the various atoms of the query structure are statistically interpreted and reported.

Because CAS defines in the connection tables nine different types of bonds (chain-single, chain-double, chain-triple, chain-tautomer, ring-single, ring-double, ring-triple, ring-tautomer, and ring-alternating), and there is a

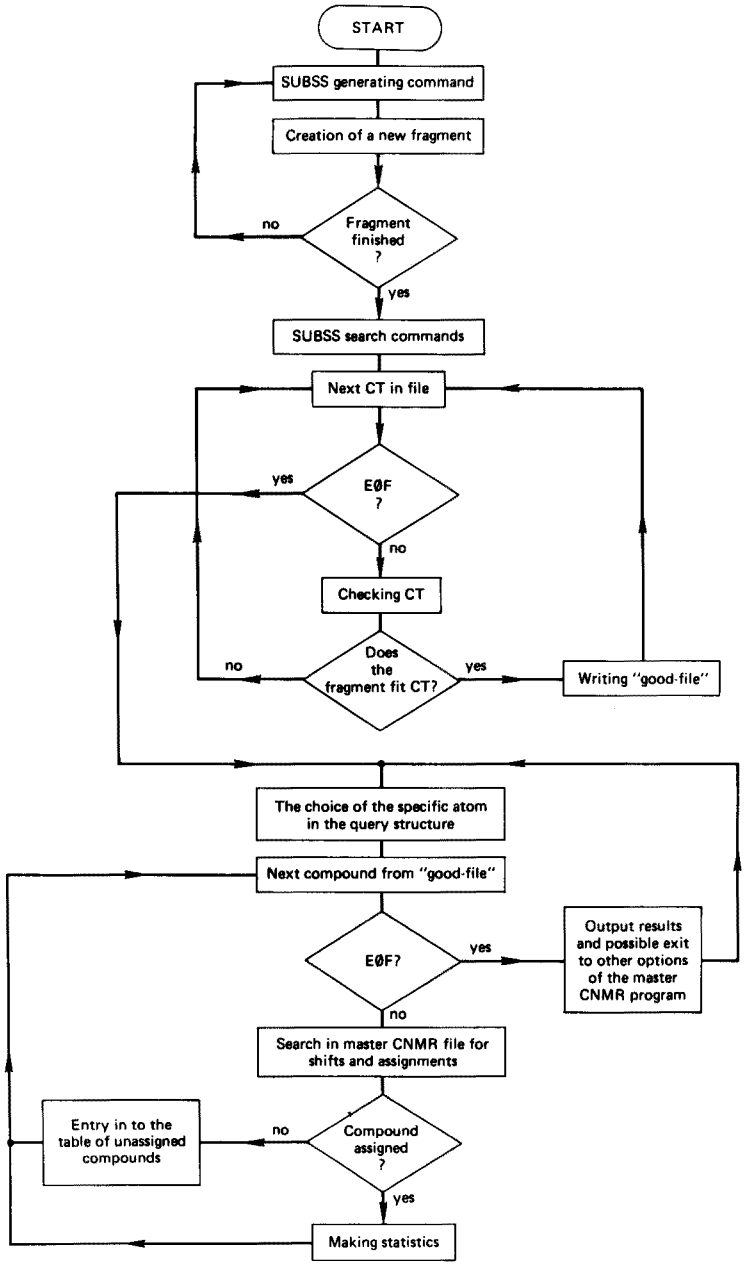


Fig. 2. Program flowchart for the described program. CT is the abbreviation for the "Connection Table".

great variety of construction commands, it is possible to build up any type of structural fragment. When the searching is complete, the output routines permit the inspection of the shift of any atom that was in the query structure. The retrieved shifts and histograms of their distributions can also be printed, as can the list of unassigned compounds containing the same structural fragment. These can be obtained on request if further assignments are planned. An example of fragment construction and the corresponding output possibilities is shown in Table 2.

Although the collection of ^{13}C-n.m.r. data is relatively small, the amount of data to be searched is large, because the average connection table consists of a 10 × 3 matrix and on average, 7 chemical shifts/assignments per compound have to be inspected. Great care has to be taken to optimize the algorithms as well as the input/output operations, which use random access files. The program is written completely in DEC-10 FORTRAN and is incorporated in the ^{13}C-n.m.r. Search System [4, 5] developed for the NIH-EPA Chemical Information System [6], which runs on DEC-10 computers. The on-line handling of data enables the users to obtain the results very quickly. The commands in the Substructure Search System are largely self-explanatory and the system has many on-line HELP messages. As a result, the learning period for most new users is very short, and the program can be used efficiently after very few trials.

RESULTS AND DISCUSSION

The first part of Table 2 shows the way in which query structures may be built. A structure can be added on the right of the Table for clarity: it appears in the actual on-line session only if requested by the option "D", for "display". The abbreviations TC and CS stand for "Tautomer Chain" and "Chain Single", respectively. The correct mnemonics for bond types can be retrieved by typing "H" after the SBOND command. Both options FPROB (fragment probe) and SUBSS 1 (substructure search on file 1) are necessary to obtain the desired results from the files. In principle, the command SUBSS could be issued alone, without a previous FPROB, but it is very time-consuming because it conducts a bond-by-bond and atom-by-atom comparison for each structure. In practice, the System will not entertain a SUBSS command on the whole file; SUBSS can only be used with respect to a temporary file such as those that are generated by searches such as FPROB. Computer time is saved by prior use of the FPROB command. This causes the program to search for atom-centered fragments and forms a temporary (and smaller) file of candidates for SUBSS, whose work is thus reduced by at least one order of magnitude. After the substructure search has been done, the compounds found are stored in a permanent file that can be used after the user exits the substructure search.

The next step begins when the CNMR search program is called and SUB (substructure) option is chosen. The program asks the number of the atom

TABLE 2

A complete example of on-line searching for the shifts of R—COOH; responses are underlined. The text shown differs slightly from the actual computer output. Some intermediate printouts have been omitted for the sake of clarity

OPTION? CHAIN 4	1??2??3??4
OPTION? ABRAN 1 AT 3	1??2??3??4 ? ? 5
OPTION? SATOM 4 5	1??2??3??40 ?
SPECIFY ELEMENT SYMBOL = O	? 50
OPTION? SBOND 3 4 3 5	1??2??3%%40 %
BOND TYPE (H FOR HELP) = TC	% 50
OPTION? SBOND 2 3	1??2**3%%40 %
BOND TYPE (H FOR HELP) = CS	% 50

OPTION? FPROB
FRAGMENT:
1C?????2C*****3C

THIS FRAGMENT OCCURS IN 1794 COMPOUNDS

FRAGMENT:
40%%%%%3C%%%%%50
*
*
*
2C

THIS FRAGMENT OCCURS IN 259 COMPOUNDS

FILE = 1, 197 COMPOUNDS CONTAIN ALL 2 FRAGMENTS

OPTION? SUBSS 1

FILE = 2, SUCCESSFUL SUBSTRUCTURES = 194

OPTION? OUT

TABLE 2 (continued)

```
.CNMR

Option: SUB

Type the number of atom you want CNMR shifts for: 3

There were  194 compounds with a given atom.
With assignments:  134, not or partial assigned:  60
Among unassigned  5 without the requested shift
Which ones do you want to see?
Type OUT to exit, otherwise A(ssigned) or U(nassigned): A

#   5: CAS reg.      57114 Shift 180.4 ppm
#   7: CAS reg.      60333 Shift 180.2 ppm
#   8: CAS reg.      64197 Shift 178.1 ppm
    •                  •       •
    •                  •       •
    •                  •       •
# 118: CAS reg.   59331952 Shift 172.1 ppm
# 119: CAS reg.   59331963 Shift 174.8 ppm

Statistics for   134 shifts
Main value:      173.4 ppm
st. deviat.:       0.6 ppm
min. value:      161.0 ppm
max. value:      185.7 ppm

Type N if you don't want the histogram [Y1]: Y3

162 (ppm), freq.: 10  **********
165 (ppm), freq.: 23  ***********************
168 (ppm), freq.: 12  ************
171 (ppm), freq.: 17  *****************
174 (ppm), freq.: 22  **********************
177 (ppm), freq.: 16  ****************
180 (ppm), freq.: 19  *******************
183 (ppm), freq.: 13  *************
186 (ppm), freq.:  2  **

In order to:              Type:

See the spectrum      – REG to find ID#, then SPEC
Assign some spectrum  – REG to find ID#, the ASS
Exit                  – OUT
Continue              – (CR):
```

of interest. The number entered must be the same as was used in the query structure. The output of all shifts considered in the calculation might of course be omitted, but it is often very useful, especially in a first pass, when misassignments may be corrected. The histogram of the shift distribution can be omitted as well, but if it is printed out, the sampling width can be

given as shown, with the command Y3 (for width 3 ppm). From the histogram shown in Table 2, it can immediately be seen that the distribution is not uniform and that the mean shift does not reflect the distribution. With additional options, such as the Registry number search and the display of structures from the ^{13}C-n.m.r. Data Base, it is very easy to determine the cause of any abnormal distribution of shifts. At this point, a new and more precise augmented substructure can be generated, and the entire run can be repeated in order to obtain a new distribution of chemical shifts.

Generally, an experienced user would try to avoid this kind of situation and begin by defining the substructure fragment more precisely. In the example shown in Table 2, the bond between the atoms 1 and 2 was not specified and this forces the SUBSS search to accept all types of bonds, aliphatic or aromatic, as possible matches to the query structure.

In order to show the selective power of this program, the example from Table 2 was worked out in more detail. The same basic fragment R—COOH was generated 3 times with different second neighbors: the resulting query structures were C—CH2—COOH, C=CH—COOH, and C=C(R)—COOH. In the second and third examples, the commands TERMA 2,2 and TERMA 3,2 were used to prevent additional substitution at atoms 2 and 3 respectively. In the fourth example, such a command was obviously unnecessary. The first histogram in Fig. 3 is the same as in Table 2, except that a different

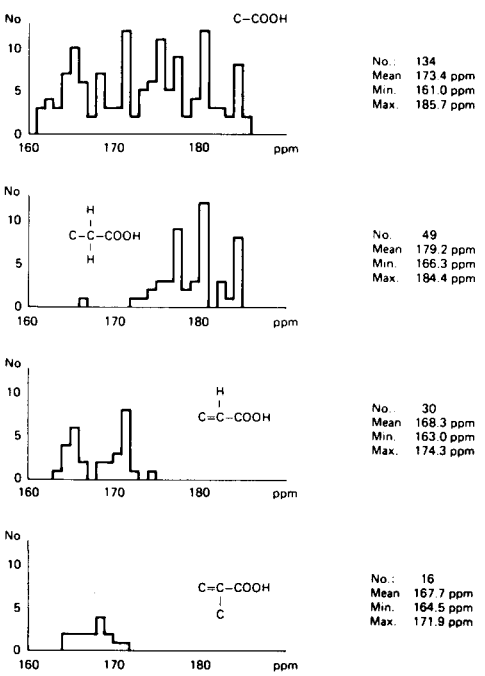


Fig. 3. The chemical shift distributions for the same central carbonyl atom with different second-neighbor environments.

sampling interval was used. The corresponding histograms of the three extended fragments containing the same central part, are shown. The contribution to the shift distribution of the specific subgroups in the first histogram becomes instantly apparent. The influence of the double bond between the first and second neighbors is responsible for the shifts in the region below 175 ppm, while the saturated carboxylic acids have shifts that, on average, are about 10 ppm lower downfield. It is also clear from these results that unless the distribution is really "gaussian-like", the standard deviations and upper or lower limits of the intervals are rather poor descriptors of the shifts. The difference in the number of compounds in the first histogram and the sum of those in the other three arises because all carboxylic acids, including those containing non-carbon atoms as the second neighbors, are considered in the first example, while in the other searches, only carbon is permitted as a substituent on the non-carboxyl carbon. It is noteworthy also that the substructure search has the command INCLA n, that makes it possible to define the possible alternative atoms to be considered as the neighbors on the same place. This command is very useful for further investigation of problems such as that in the example presented.

The most serious shortcoming of this system is related to the number of compounds in the file. At this stage, with about 4,000 spectra in hand, it is rather unrealistic to expect good results when more than second-order neighbors are included, although in some cases, such as the second example in Fig. 3, this might still be valuable. In any event, as the data base increases in size this shortcoming will become less apparent.

One of us (J. Z.) acknowledges the partial support of this work by the Research Community of Slovenia.

REFERENCES

1 F. W. Wehrli and T. Wirthlin, Interpretation of Carbon-13 NMR Spectra, Heyden, London, 1976.
2 E. Pretsch, J. T. Clerc, J. Seibl, and W. Simon, Tabellen zur Strukturaufklärung organischer Verbindungen, Springer-Verlag, Berlin, New York, 1976, pp. B5—B10.
3 J. B. Stothers, Carbon-13 NMR Spectroscopy, Academic Press, New York, 1972.
4 CNMR Data Base, NIC, Delft, Holland (Attn: C. Citroen, NIC, CID-NTO, PO Box 36 2600 AA, Delft, The Netherlands).
5 D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne and S. R. Heller, Org. Mag. Res., 11 (1978) 000.
6 S. R. Heller, G. W. A. Milne, and R. J. Feldmann, Science, 253 (1977) 195.
7 R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller and B. Koch, J. Chem. Inf. Comp. Sci., 17 (1977) 157.
8 J. A. Miller, Substructure Search System, Users Manual, Fein-Marquart Associates, Inc., 7215 York Road, Baltimore, Md., 21212.
9 D. E. Knuth, The Art of Computer Programming, 2nd edn., Addison-Wesley, Reading, 1973, Vol. III, p. 125.

Analytica Chimica Acta, 103 (1978) 151—165
Computer Techniques and Optimization

THE LEARNING MACHINE IN QUANTITATIVE CHEMICAL ANALYSIS
Part 1. Anodic Stripping Voltammetry of Cadmium, Lead and Thallium

M. BOS* and G. JASINK

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

(Received 6th March 1978)

SUMMARY

The linear learning machine method was applied to the determination of cadmium, lead and thallium down to 10^{-8} M by anodic stripping voltammetry at a hanging mercury drop electrode. With a total of three trained multicategory classifiers, concentrations of Cd, Pb and Tl could be predicted with an accuracy of ±10%. The classifiers were trained with the use of least-squares minimization. Numerical problems in the data matrix inversion were overcome by using singular value decomposition.

Pattern recognition has been applied in chemical analysis in various fields, i.e. infrared spectroscopy [1], mass spectrometry [2], nuclear magnetic resonance [3] and stationary electrode polarography [4]. Most of the applications described so far refer to qualitative analytical aspects such as identification problems, the presence or absence of a second component, multiplicity of peaks, etc. Some quantitative problems with a restricted number of classes were solved by the use of sets of binary classifiers e.g. in the number of carbon atoms in a given compound from its mass spectrum [5]. When a concentration has to be classified, this method becomes impractical because of the large number of classifiers needed if a reasonable accuracy is required over even a moderate concentration range.

Anodic stripping voltammetry (a.s.v.) is a very sensitive technique which is used extensively for determination of the heavy metal content of natural waters, body fluids, etc. Although the differential pulse mode makes this technique quite selective, cases with badly overlapping peaks remain in which classical evaluation of the results is impossible. Peak deconvolution could be used to resolve these problems, but the process is time-consuming and requires a knowledge of what to look for, otherwise serious errors will occur. Pattern recognition is fast — at least after training is complete — and requires no knowledge about the unknown samples other than that they should resemble samples of the training set. Moreover, in the fields where this type of analysis is being used, often only a go/no go decision is required from the result of the analysis.

Accordingly, the learning machine method developed by Kowalski et al. [6] where the classifier operating directly on the data produces the quantitative measure of interest, is the best to apply. As Tunnicliff and Wadsworth [7] have pointed out, numerical problems can arise in handling the data matrix. In the present work, singular value decomposition of the data matrix and a modified Ho—Kashyap algorithm [8] were used to circumvent this problem.

THEORY

A short review on the theory of linear learning machines is given here. Anodic stripping voltammograms, like any other set of measurements, can be represented as patterns $X1, X2, \ldots, Xi, \ldots, Xd$, where, for a.s.v., Xi denotes the current measured at point i on the voltage axis and d is the number of measurements taken along that axis. The vector X (pattern $X1, X2, \ldots, Xd$) can be thought of as a point in the d-dimensional euclidian vector space E_d. A given phenomenon is associated with pattern X. This phenomenon belongs to one out of R classes. The learning machine problem is to find from a set of patterns with known classification (the training set) a classifier (or a set of classifiers) that correctly classifies an unknown pattern. The euclidian space E_d can be divided into R parts, each of which corresponds to one of the R classes. Classes are separated by so-called decision planes. These planes are defined implicitly by a set of functions $g1\ (X), g2\ (X), \ldots, gR(X)$ of the pattern vector X. The decision functions are chosen in such a manner that if X belongs to class i, then

$$gi\ (\mathrm{X}) > gj\ (\mathrm{X}) \text{ for } j = 1, 2, \ldots, R \text{ with } j \neq i \tag{1}$$

If the functions $g1(\mathrm{X}), g2(\mathrm{X}), \ldots, gR(\mathrm{X})$ are known, the pattern X can be classified by determining the maximum value. A simple case is that of only two classes ($R = 2$), in which a single decision function suffices:

$$g(\mathrm{X}) = g1(\mathrm{X}) - g2(\mathrm{X}) \tag{2}$$

if $g(\mathrm{X}) > 0$ then X belongs to class 1, and if $g(\mathrm{X}) < 0$ then X belongs to class 2. The decision surface can then be represented by

$$g(\mathrm{X}) = 0 \tag{3}$$

In the linear learning machine, the decision plane is given by

$$g(\mathrm{X}) = W_1X_1 + W_2X_2 + \ldots + W_dX_d + W_{d+1} = 0 \tag{4}$$

This is a hyperplane with the dimension $(d - 1)$ separating hyperspace E_d into two parts.

During the training, the parameters $W_1, W_2, \ldots W_d, W_{d+1}$ are adjusted in such a way that the hyperplane $g(\mathrm{X}) = 0$ separates the pattern points of the training set correctly. The hyperplane $g(\mathrm{X}) = 0$ does not pass through the origin of the space E_d. As it is easier to work with a decision plane through the origin, a constant element c is added to the vector X:

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d, Y_{d+1}) = (X_1, X_2, \ldots, X_d, c) = (\mathbf{X}, c) \tag{5}$$

Then $g(\mathbf{X})$ can be replaced by

$$g(\mathbf{y}) = W_1 Y_1 + W_2 Y_2 + \ldots + W_d Y_d + W_{d+1} Y_{d+1} \tag{6}$$

which can be written as

$$g(\mathbf{Y}) = (\mathbf{W}.\mathbf{Y}) \tag{7}$$

In the following treatment, $n = d + 1$ and the vectors W and Y are vectors in euclidian space E_n. If the sign of the patterns belonging to class 2 is changed, eqns. (2) and (7) can be combined:

$$(\mathbf{W}.\mathbf{Y}_j) > 0 \quad j = 1, 2, \ldots, m \tag{8}$$

with m the number of patterns in the training set. This inequality can be replaced [8] by $(\mathbf{W}.\mathbf{Y}_j) = b_j$ with $b_j > 0$ and $j = 1, 2, \ldots, m$, which can be rewritten as

$$A\mathbf{W} = \mathbf{b} \tag{9}$$

Here A is the data matrix $(m \times n)$ of which the rows are constituted by the patterns $\mathbf{Y}_j, j = 1, 2, \ldots, m$.

Least-squares training

In general, $m > n$, so that the set of equations (9) is overdetermined and has no exact solution. If the training set is linearly separable, $\mathbf{b} > 0$ (notation $\mathbf{b} > 0$ means $b_j > 0, j = 1, 2, \ldots, m$) can be determined with the Ho—Kashyap [8] algorithm in such a way that a value of W can be found which satisfies $A.\mathbf{W} > 0$. This algorithm minimizes the function

$$J(\mathbf{W},\mathbf{b}) = \|A\mathbf{w} - \mathbf{b}\|^2 \tag{10}$$

for w and for b, hence the name least-squares training. The algorithm is iterative with k the number of the iteration step being processed: for $\mathbf{b}^{(0)} > 0$, with initial arbitrary values of b,

$$\mathbf{W}^{(k)} = A^{+}\mathbf{b}^{(k)} \tag{11}$$

In this equation A^{+} is a generalized inverse of A:

$$\mathbf{e}^{(k)} = A\mathbf{w}^{(k)} - \mathbf{b}^{(k)} \tag{12}$$

if $e_j^{(k)} > 0$, then $b_j^{(k+1)} = b_j^{(k)} + 2\rho e_j^{(k)}$ with $0 < \rho < 1$

$e_j^{(k)} < 0$, then $b_j^{(k+1)} = b_j^{(k)}$ (13)

The sequence (11) through (13) is repeated until

$$A\mathbf{w}^{(k)} > 0 \tag{14}$$

This algorithm has the following properties: if $e_i^{(k)} < 0$ for all j, $1 < j < m$, then the training set is not linearly separable; if $e_j^{(k)} > 0$ for all j, $1 < j < m$ and for $k = 1, 2, \ldots, k$, and eqn. (14) is not yet satisfied, then there may be

a solution. If this solution exists, it will be found in a finite number of iterations. If it does not exist, then $\mathbf{e}^{(k)}$ approaches a limiting vector.

Least-squares training for (nearly) dependent columns in data matrix

The calculation of the generalized inverse A^+ from the data matrix A requires special attention. If the columns of A are linearly independent, A^+ can be replaced by $A^+ = (A^t.A)^{-1}.A^t$

If A is ranked much lower than n, the matrix $(A^t.A)$ is singular and A^+ cannot be calculated from this equation.

To restate the problem: the set of m equations with n unknowns of eqn. (9), $A\mathbf{w} = \mathbf{b}$, must be solved for $m > n$. Matrix A has n columns each consisting of m elements. These columns of A can be represented as the (column) vectors a1, a2, ..., $\mathbf{a}_n$ in space E_m, which is the m-dimensional euclidian vector space. The vector b also lies in E_m. The vector W has n elements $(W_1, W_2, \ldots, W_n)$ and can be represented in E_n.

The n columns of the matrix A span a subspace of E_m. This subspace, denoted as $R(A)$, contains all vectors that are a linear combination of the columns of A. For all arbitrary values of α_i, $i = 1, 2, \ldots, n$, the vector d defined as $\mathbf{d} = \alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots\ldots + \alpha_n\mathbf{a}_n$ lies in $R(A)$. The vector d defined as $\mathbf{d} = W_1\mathbf{a}_1 + W_2\mathbf{a}_2 + \ldots. + W_n\mathbf{a}_n$, therefore also lies in $R(A)$ for all possible values of W_i, $i = 1, 2, \ldots., n$. Thus for all vectors w from E_n, d lies in $R(A)$. If the vector b from E_m (cf. eqn. 9) lies in $R(A)$, then there is a vector w from E_n for which $\mathbf{d} = A\mathbf{w}$ coincides with b and eqn. (9) can be satisfied.

If the vector b is not contained in $R(A)$, then there is no vector w from E_n for which $\mathbf{d} = A\mathbf{w}$ coincides with b. For all vectors w from E_n, there is then a difference vector: $\mathbf{e} = \mathbf{d} - \mathbf{b} = A\mathbf{w} - \mathbf{b}$, which is different from the null vector. The distance between d and b is the length of the difference vector:

$$\|\mathbf{d} - \mathbf{b}\| = \|\mathbf{e}\| = \left(\sum_{j=1}^{m} e_j^2\right)^{\frac{1}{2}} \tag{15}$$

For the solution w in E_n of $A\mathbf{w} = \mathbf{b}$ satisfying the least-squares criterion, the distance between d and b is minimized:

$$\|\mathbf{e}\|^2 = \|A\mathbf{w} - \mathbf{b}\|^2 = J(\mathbf{w}) \tag{16}$$

The distance between d and b is minimal when d is the normal projection of b onto the subspace $R(A)$. The normal projection of a point (b) from the space E_m onto the subspace $R(A)$ of E_m is a unique point of the subspace $R(A)$ and is denoted as d(proj). If the point d(proj) of $R(A)$ has been found, then vector w can be found, satisfying $A\mathbf{w} = \mathbf{d}(\text{proj})$.

The matrix A contains n columns. If only r of these columns are linearly independent, then the other $(n - r)$ columns are linear combinations of the first r columns. In that case, the rank of the matrix A is r and the set of equations $A\mathbf{w} = 0$ has $(n - r)$ independent solutions $\mathbf{w}_1, \mathbf{w}_2, \ldots., \mathbf{w}_{n-r}$. If w(spec) is a solution of $A\mathbf{w} = \mathbf{d}(\text{proj})$, the general solution of $A\mathbf{w} = \mathbf{d}(\text{proj})$ becomes

$$\mathbf{w} = \mathbf{w}(\text{spec}) + \sum_{i=1}^{n-r} \alpha_i \mathbf{w}_i \tag{17}$$

with α_i arbitrary. For one of these solutions w of eqn. (17), $\|\mathbf{w}\|$ is minimal. This solution can be found with a pseudo-inverse [9, 10]. If the rank of the matrix A equals r, the subspace $R(A)$ of E_m can be spanned by only r columns of A. If the columns are aligned in such a way that $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ are linearly independent, then $\mathbf{a}_{r+1}, \mathbf{a}_{r+2}, \ldots, \mathbf{a}_n$ are linearly dependent on $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ with $a_{r+1} = \Sigma_{j=1}^{r} \alpha_j \cdot \mathbf{a}_j$ and $\mathbf{a}_n = \Sigma_{j=1}^{r} w_j \mathbf{a}_j$. The vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ now form a base of the subspace $R(A)$; the dimension of the subspace is r.

Singular value decomposition to determine the rank of the data matrix

To determine the orthonormal projection d(proj) of vector b from E_m onto $R(A)$, a base for $R(A)$ must be known as well as the rank of A or the dimension of $R(A)$. This problem can be solved by calculation of the pseudo-inverse of A based on singular value decomposition (SVD) [9—14]. The pseudo-inverse of the $m \times n$ matrix A is the $n \times m$ matrix A^+ with the property that

$$\mathbf{w} = A^+\mathbf{b} \tag{18}$$

is a solution of $A\mathbf{w} = \mathbf{d}(\text{proj})$, for which $\|\mathbf{w}\|$ is minimal if $A\mathbf{w} = \mathbf{d}(\text{proj})$ has more than one solution. In the SVD, an arbitrary matrix A can be decomposed to 3 matrices with special properties:

$$A = U\Sigma V^t \tag{19}$$

in which A is the $m \times n$ arbitrary matrix ($m \geqslant n$); U is the $m \times n$ orthonormal matrix (for columns), i.e. $U^tU = I_n$; V is the $n \times n$ orthonormal matrix (for columns and rows), i.e. $V^tV = I_n$ and $VV^t = I_n$; and Σ is the $n \times n$ diagonal matrix with diagonal elements $\delta_1 \geqslant \delta_2 \geqslant \delta_3 \geqslant \ldots > \delta_n \geqslant 0$, the other elements being zero. The elements δ_i of the diagonal matrix Σ are called the singular values of A.

The pseudo-inverse of matrix A is

$$A^+ = V\Sigma^+U^t \tag{20}$$

in which Σ^+ is the $n \times n$ diagonal matrix with elements $\delta_1^+, \delta_2^+, \ldots, \delta_n^+$ with $\delta_i^+ = 1/\delta_i$ for $\delta_i > 0$, and $\delta_i^+ = 0$ for $\delta_i = 0$ [9, 10].

In theory, the rank of A equals the number of the singular values greater than zero. If this number is r, then a base for the subspace $R(A)$ is formed by the first r columns of matrix U. These are the only columns of U that are of importance in eqns. (19) and (20); the other $(n - r)$ columns of U are unused because the corresponding $(n - r)$ elements of Σ are zero.

In practice, A can contain a large number of nearly dependent columns. In this case, Σ has a number of elements δ_i that are very small. These are the higher numbered δ_i values in the above $\delta_1 \ldots \delta_n$ sequence. If the first q singular values are relatively much larger than the others, A^+ can be calculated

from eqn. (20) with Σ^+ replaced by $\Sigma^+_{(q)}$, which is the $n \times n$ diagonal matrix with elements

$$\delta_1^+ = 1/\delta_1, \delta_2^+ = 1/\delta_2, \ldots, \delta_q^+ = 1/\delta_q \text{ and } \delta_{q+1} = \delta_{q+2} = \ldots = \delta_r = \ldots = \delta_n = 0 \qquad (21)$$

In fact, matrix A has been replaced by matrix $A(q)$ with rank q. The base of space $R(A_{(q)})$ is then formed by the first q columns of U. In this way, the value of the "norm" of the difference $A - A(q)$

$$\|A - A(q)\| = \left(\sum_{i=q+1}^{n} \delta_i^2 \right)^{\frac{1}{2}} \qquad (22)$$

is smallest for a given q value (see [10]). $A\mathbf{w} = \mathbf{b}$ can then be solved by means of eqns. (20) and (21).

The solution $\mathbf{w}_q$ *with* q *non-zero singular values*

The classifier $\mathbf{w}$ that is required has to satisfy only $\mathbf{d} = A\mathbf{w} > 0$. SVD gives $A = U\Sigma V^t$, of which the elements are

$$a_{ik} = \sum_{l=1}^{n} U_{il} V_{kl} \delta_l \qquad (23)$$

Here a_{ik} is the i-th row and k-th column element of the matrix A. U_{il} and V_{kl} are the elements of U and V defined in the same manner. From eqn. (23):

$$\mathbf{d} = A\mathbf{w} = \sum_{k=1}^{n} \delta_k (\mathbf{V}_k . \mathbf{w}) \mathbf{u}_k \qquad (24)$$

where $(\mathbf{v}_k . \mathbf{w})$ represents the inner product of the k-th column of V and the vector $\mathbf{w}$. This inner product is defined in E_n. From eqn. (24), it follows that $\mathbf{d}$ is a linear combination of the columns $\mathbf{u}_k$ of U. If the rank of A equals r, this can be seen from $\delta_{k=0}$ for $k > r$. Equation (24) then reduces to

$$\mathbf{d} = \sum_{k=1}^{r} \delta_k (\mathbf{v}_k . \mathbf{w}) \mathbf{u}_k \qquad (25)$$

For $\mathbf{w} = A^+\mathbf{b}$ (eqn. 18), an expression in the elements of U, V and Σ can also be found (cf. eqn. 20).

If the rank of A equals r, then $\delta_i^+ = 1/\delta_i$ for $i \leqslant r$ and $\delta_i^+ = 0$ for $i > r$. From eqns. (18), then

$$\mathbf{w} = \sum_{i=1}^{r} (1/\delta_i) (\mathbf{u}_i . \mathbf{b}) \mathbf{v}_i \qquad (26)$$

where $(\mathbf{u}_i . \mathbf{b})$ is the inner product of the i-th column of U and the vector $\mathbf{b}$, defined in E_m. The vectors $\mathbf{d}$ (cf. eqn. 25) and $\mathbf{w}$ (cf. eqn. 26) correspond to $\mathbf{w} = A^+\mathbf{b}$ and $\mathbf{d} = A.\mathbf{w}$ calculated with the full rank-r pseudo-inverse A^+.

If $\mathbf{w}$ is calculated with only q terms ($q < r$), then

$$\mathbf{w}_q = \sum_{i=1}^{q} (1/\delta_i)\,(\mathbf{u}_i.\mathbf{b})\,\mathbf{v}_i \tag{27}$$

This equation indicates that $\mathbf{w}_q$ is a linear combination of the columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ of V. V is orthonormal (cf. $V^tV = I_n$ above), and so for the columns of V, $(\mathbf{v}_k.\mathbf{v}_i) = 1$(for $k = i$) or 0 for $k \neq i$.

The subspace of E_n spanned by the vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q)$ is the orthogonal complement of the subspace spanned by $(\mathbf{v}_{q+1}, \mathbf{v}_{q+2}, \ldots, \mathbf{v}_n)$. A vector from the one subspace is orthogonal to all vectors from the other orthogonally complementary subspace. Thus $\mathbf{w}_q$ is orthogonal to $(\mathbf{v}_{q+1}, \mathbf{v}_{q+2}, \ldots, \mathbf{v}_n)$; i.e. $(\mathbf{v}_k.\mathbf{w}_q)$ is an arbitrary value for $k = 1, 2, \ldots, q$, or is zero for $k = q + 1, q + 2, \ldots, n$. Thus

$$\mathbf{d}\,(\mathbf{w}_q) = A\mathbf{w}_q = \sum_{k=1}^{q} \delta_k\,(\mathbf{v}_k.\mathbf{w}_q)\,\mathbf{u}_k + \sum_{k=q+1}^{r} \delta_k\, o\mathbf{u}_k \tag{28}$$

or

$$\mathbf{d}\,(\mathbf{w}_q) = \sum_{k=1}^{q} \delta_k\,(\mathbf{v}_k.\mathbf{w}_q)\,\mathbf{u}_k$$

where $\mathbf{d}\,(\mathbf{w}_q)$ is the vector $\mathbf{d}$ corresponding to a $\mathbf{w}$ value calculated from q terms only. Equation (27) gives

$$(\mathbf{v}_k.\mathbf{w}_q) = \left(\mathbf{v}_k.\left(\sum_{i=1}^{q} \frac{1}{\delta_i}(\mathbf{u}_i.\mathbf{b})\,\mathbf{v}_i\right)\right) = \sum_{i=1}^{q} \frac{1}{\delta_i}(\mathbf{u}_i.\mathbf{b})\,(\mathbf{v}_k.\mathbf{v}_i) \tag{29}$$

Combination of eqn. (29) with the term $(\mathbf{v}_k.\mathbf{v}_i) = 1$(for $k = i$) gives

$$(\mathbf{v}_k.\mathbf{w}_q) = \frac{1}{\delta_k}\,(\mathbf{u}_k.\mathbf{b}) \tag{30}$$

and eqns. (28) and (30) finally give

$$\mathbf{d}\,(\mathbf{w}_q) = \sum_{k=1}^{q} (\mathbf{u}_k.\mathbf{b})\,\mathbf{u}_k \tag{31}$$

Thus for the calculation of $\mathbf{d}\,(\mathbf{w}_q)$, it is not necessary to calculate $\mathbf{w}_q$.

For a classifier $\mathbf{w}$, it is sufficient to satisfy $\mathbf{d} = A\mathbf{w} > 0$. With eqn. (31), a vector $\mathbf{b}$ can be found for a specific choice of q in such a way that $\mathbf{d}\,(\mathbf{w}_q) > 0$, so that $A\mathbf{w}_q > 0$ is also then valid. With this q value and this $\mathbf{b}$ vector, $\mathbf{w}_q$ can be calculated from eqn. (27).

In the search for a vector $\mathbf{b}$ to satisfy $\mathbf{d}\,(\mathbf{w}_q) > 0$, a modified Ho—Kashyap algorithm can be used: for arbitrary positive initial values of $\mathbf{b}$, i.e. $\mathbf{b}^{(0)} > 0$,

$$\mathbf{d}^{(k)} = \sum_{i=1}^{q} (\mathbf{u}_i.\mathbf{b}^{(k)})\,\mathbf{u}_i \tag{32}$$

$$\mathbf{e}^{(k)} = \mathbf{d}^{(k)} - \mathbf{b}^{(k)} \tag{33}$$

$$\text{If } e_j^{(k)} > 0, \text{ then } b_j^{(k+1)} = b_j^{(k)} + 2\rho e_j^{(k)} \text{ with } 0 < \rho < 1 \tag{34}$$

$$\text{If } e_j^{(k)} \leqslant 0, \text{ then } b_j^{(k+1)} = b_j^{(k)}$$

The calculations based on the relationships (32)—(34) must be repeated until $\mathbf{d}^{(k)} > 0$.

The properties of this modified Ho—Kashyap training algorithm regarding termination in the case of a linearly inseparable training set are the same as for the original algorithm.

U is orthonormal, i.e. $U^+U = I_n$, thus for the U columns, $(\mathbf{u}_k.\mathbf{u}_i) = 1$ for $k = i$, or 0 for $k \neq i$. The subspace spanned by the columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ is denoted as $R(U_q)$. The orthogonal projection of the vector $\mathbf{b}$ from E_m onto the subspace $R(U_q)$ of E_m is given by

$$\mathbf{b}' = \sum_{k=1}^{q} (\mathbf{u}_k.\mathbf{b})\, \mathbf{u}_k \tag{35}$$

Equation (35) is identical with eqn. (31), and so $\mathbf{d}\ (\mathbf{w}_q)$ is the projection of $\mathbf{b}$ onto $R(U_q)$.

Selection of the number of non-zero singular values, q

If the measurements comprising the data matrix A were without experimental error, the dimensionality of the subspace $R(A)$ would be low, which can be denoted by q(ideal). Young and Calvert [15] called this the intrinsic dimensionality of the training set. Because of random errors during measurements, the mathematical rank of the matrix A equals r. Thus $q(\text{ideal}) < r \leqslant n$.

The low numbered columns of matrix U form the most important vectors spanning $R(A)$. It is likely that the influence of the phenomena to be classified on the measurement will exceed the influence of random errors. Therefore the best choice for q will involve including the first q columns of U for which $\mathbf{d}\ (\mathbf{w}_q) > 0$ can be found.

Selection of the right-hand vector, b

If a binary classifier is required, the vector $\mathbf{b}$ can be initialized with a set of arbitrary positive values. The modified Ho—Kashyap training algorithm (eqns. 32—34) can then be used to adjust this $\mathbf{b}$ vector until $\mathbf{d}^{(k)} > 0$; $\mathbf{w}_q$ can then be calculated.

If a multicategory classifier is required, the vector $\mathbf{b}$ is initialized with a measure of the property of the patterns. In the a.s.v. experiments, this can be done with a concentration measure for the compound of interest. Then $\mathbf{w}$ must be calculated directly from eqn. (27) with the lowest possible q value, giving a small value for

$$\|e\| = \|A\mathbf{w} - \mathbf{b}\| \tag{36}$$

EXPERIMENTAL

Chemicals

The supporting electrolyte used was 10^{-2} M potassium chloride (Merck, Suprapur) and 10^{-4} M hydrochloric acid (Merck, Suprapur). Ultrapure nitrogen was passed through an acidic chromium(II) solution to remove the last traces of oxygen down to 1 ppm [16], and then through an all-glass tubing system to the cell to remove oxygen from the solution; purging was continued for 15 min. The mercury for the hanging mercury drop electrode (HMDE) was purified by washing it with nitric acid and vacuum-distilling it twice. The 10^{-6}—10^{-8} M solutions of lead, thallium and cadmium were prepared daily from 0.1 M stock solutions of the nitrates (Merck, reagent grade).

Apparatus

A PAR Model 174 polarograph was connected to a PDP-11/10 (Digital Equipment Corp.) online computer. The computer samples the cell current available at the recorder terminals of the PAR 174 with a 12-bit A/D-converter (range ±5 V) contained in the Laboratory Peripheral System (Digital Equipment Corp.).

The electrodes were a HMDE (Radiometer P958 b), a saturated calomel reference electrode, and a platinum wire auxiliary electrode; the auxiliary electrode was connected to the measuring vessel by an agar—potassium chloride salt bridge. A Radiometer type SMP-1 stirring motor was used with a glass stirring rod. The cell was thermostatted to 20 ± 0.5 °C.

All glassware was degreased with dichromate/sulphuric acid mixture, and washed with (1 + 1) nitric acid and doubly distilled water. When not in use, the glassware was kept in doubly distilled water.

Procedure

Measurements were done in the differential pulse mode at a scan rate of 5 mV s^{-1}, with a measuring interval of 0.5 s and a pulse height of 25 mV. A typical run was performed as follows. Purified nitrogen was passed through the solution for 15 min. The metals were then concentrated on the HMDE by applying —0.9 V to the cell, with stirring, for 3 min. Stirring was then stopped, and after 30 s the mode was switched to differential pulse; after another 30 s, the anodic scan was started. The PDP-11/10 automatically senses the start of the scan and with a delay of 180 mV acquires the current data at 480-ms intervals in the range of —0.72 to —0.22 V.

Computer programs

Data acquisition program. The program for data acquisition is straightforward and runs on the PDP-11/10 online computer. The data acquired during the runs can be stored on disks, together with pertinent information. The information stored can be typed and punched in ASCII format for transference to another computer.

Training program. The training of the learning machine was done on a DEC-10 computer. Input to the training program was the ASCII paper-tape containing the results of the measurements in the form of integers between 2048 and 4096 (A/D-converter format for voltages between 0 and +5 V). These values are converted to real format and adjusted in accordance with the position of the current range switch on the PAR 174 used during the experiment. The dynamic range of the patterns becomes 42—420000. For the training of binary classifiers, the concentration range for a specific compound is divided into a number of intervals, and classifiers are trained that classify the pattern on the correct sides of the boundaries between intervals, by means of the singular value decomposition and modified Ho—Kashyap algorithm described above.

The singular value decomposition was done with the use of a subroutine from the Numerical Algorithms Group (Oxford, England). This routine is based on the method of Golub and Reinsch [17]. In calculations of the multicategory classifiers, the elements b_j of the right-hand vector are given the value of the concentrations of the metal ion of interest as the initial value. Matrix U is not adjusted in this case. In fact, the program for this case is a simplification of the former one in which calculation is stopped after the zero-th iteration of the modified Ho—Kashyap algorithm.

Program for evaluation of results. To test the performance of the multicategory classifiers, the products (**w**.**y**) must be calculated. The last element of the patterns is a constant factor c; thus the last term from (**w**.**y**) is $w_n c$.

To simplify the calculations and the presentation this last term is omitted in the table entries, but is indicated separately. Under normal conditions, a.s.v. results for a mixture equal the sum of the results for the separate compounds. If $\mathbf{y}p = (\mathbf{p},c)$ is a pattern recorded for metal ion P, and $\mathbf{y}q = (\mathbf{q},c)$ is a pattern recorded for metal ion Q, then a mixture of P and Q gives a pattern:

$$\mathbf{y}(p,q) = (\mathbf{p} + \mathbf{q},c) \tag{37}$$

If the last element of the weight vector **w** is w_n then:

$$(\mathbf{w}.\mathbf{y}p) = (\mathbf{w}.(\mathbf{p},c)) = (\bar{\mathbf{w}}.\mathbf{p}) + w_n c \tag{38}$$

$$(\mathbf{w}.\mathbf{y}q) = (\mathbf{w}.(\mathbf{q},c)) = (\bar{\mathbf{w}}.\mathbf{q}) + w_n c \tag{39}$$

$$(\mathbf{w}.\mathbf{y}p,q) = (\bar{\mathbf{w}}.(\mathbf{p} + \mathbf{q})) + w_n c = (\bar{\mathbf{w}}.\mathbf{p}) + (\bar{\mathbf{w}}.\mathbf{q}) + w_n c \tag{40}$$

Thus, classification of a mixture of P and Q can be simulated by calculating ($\bar{\mathbf{w}}$.**p**) and ($\bar{\mathbf{w}}$.**q**), determining the sum and adding $w_n c$. In this way, presentation of the results can be restricted to the value of the ($\bar{\mathbf{w}}$.**p**) for single-component samples of the various metal ions. Moreover, the value of $w_n c$ turned out to be very small in comparison to the other terms from ($\bar{\mathbf{w}}$.**p**) and could be neglected.

RESULTS

The patterns recorded by the computer contained 209 points. From these points, 49 were selected at equally spaced voltage intervals of 7.2 mV between −0.642 and −0.297 V. The training set consisted of 7 patterns for each of the three metal ions (Pb, Cd and Tl). The concentrations for these 7 patterns were 10^{-8}, 2.5×10^{-8}, 5×10^{-8}, 10^{-7}, 2.5×10^{-7}, 5×10^{-7} and 10^{-6} M; the codes used to represent these concentrations were 10, 25, 50, 100, 250, 500 and 1000, respectively. The constant c for the 50th element of the pattern was taken as 10000.

Table 1 shows the peak currents in mA for the members of the training set. Peak positions were −0.380 ± 0.005 V for Pb, −0.580 ± 0.005 V for Cd, and −0.445 ± 0.005 V for Tl.

Separate classifiers were calculated for lead, thallium and cadmium by using 2, 3, 4, 5 and 16 non-zero singular values and the 3 × 7 training set. The calculations were carried out in the following sequence: singular value decomposition of the data matrix formed by the training set, selection of the required number of non-zero singular values, and calculation of the weight vector with the use of the coded concentrations as the right-hand vector b.

Table 2 shows the recognition performance of the weight vectors for lead, cadmium and thallium obtained with 4 non-zero singular values. In these calculations, patterns y belonging to the training set were classified by calculation of $w_i.y_i$, omitting the term $w_{50}.10000$. The results are listed in the column 'Recognition with shift 0 mV'. The same calculations were performed with patterns from the computer measurements comprising the training set, in which 49 data points were selected which were shifted a fixed amount along the voltage axis with regard to the original 49 points used for training. These results are given in the columns 'Recognition with shift −5, −10, +5 and +10 mV'.

Patterns simulated by linear interpolation of the current values of the patterns comprising the training set were used to test the capability for continuous classification of the weight vectors. Table 3 shows the results of the classification of these simulated patterns.

TABLE 1

Peak currents during anodic stripping voltammetry of Pb, Cd and Tl

Concn. (M)	Peak current (mA)		
	Pb	Cd	Tl
10^{-8}	0.094	0.060	0.041
2.5×10^{-8}	0.240	0.184	0.090
5.0×10^{-8}	0.410	0.365	0.185
10^{-7}	0.847	0.790	0.353
2.5×10^{-7}	2.13	2.07	0.938
5.0×10^{-7}	4.35	4.11	1.853
10^{-6}	9.10	8.36	3.71

TABLE 2

Recognition performance of weight vectors for Cd, Pb and Tl

Sample composition ($\times 10^{-8}$ M)			Recognition with shift 0 mV			Recognition with shift −5 mV			Recognition with shift −10 mV			Recognition with shift +5 mV			Recognition with shift +10 mV		
Cd	Tl	Pb	Cd	Tl	Pb	Cd	Tl	Pb	Cd	Tl	Pb	Cd	Tl	Pb	Cd	Tl	Pb
—	—	10	0.3	1.4	10.6	0.3	1.4	10.6	0.3	1.4	10.6	0.3	1.4	10.6	0.3	1.4	10.6
—	—	25	0.2	1.3	26.8	0.3	1.2	27.5	0.2	2.0	27.5	0.2	1.4	25.6	0.2	2.1	24.0
—	—	50	0.3	0.5	46.3	0.4	0.8	47.4	0.3	2.4	47.2	0.3	1.0	44.6	0.2	2.3	41.8
—	—	100	0.8	0.4	95.2	0.9	1.3	96.6	0.8	4.4	97.0	0.7	1.2	91.7	0.6	4.1	85.2
—	—	250	0.4	−0.9	237.6	0.6	−0.3	242.4	0.6	5.6	244.3	0.2	2.3	227.2	−0.3	10.3	210.3
—	—	500	0.1	0.3	501.2	−0.3	19.9	494.8	−1.1	53.3	478.0	−0.1	−6.6	493.2	−0.7	−0.3	474.2
—	—	1000	−0.6	−0.2	1002.2	0.6	−5.9	1035.4	0.8	12.2	1050.3	−1.9	25.1	950.5	−3.4	55.8	881.9
10	—	—	7.2	1.4	1.9	7.1	1.3	2.0	6.9	1.3	2.1	7.2	1.4	1.8	7.0	1.7	1.7
25	—	—	22.0	1.5	1.8	22.0	0.9	2.0	21.3	0.6	2.1	21.5	1.9	1.5	21.0	2.7	1.3
50	—	—	43.8	1.1	2.9	43.1	0.3	3.2	41.6	−0.1	3.4	43.7	2.1	2.5	42.9	3.5	2.0
100	—	—	94.7	0.9	3.4	94.4	−0.9	4.0	92.1	−2.5	4.5	93.8	3.6	2.5	91.1	7.2	1.5
250	—	—	248.8	0.3	2.5	246.3	−5.1	4.0	239.1	−8.6	5.1	245.3	7.2	0.5	239.2	15.4	−1.6
500	—	—	489.8	−10.1	−0.2	475.6	−17.8	2.1	458.4	−23.7	4.0	492.8	0.8	−3.2	483.6	15.1	−6.7
1000	—	—	1005.8	4.5	−1.6	1001.6	−18.6	5.0	973.9	−35.3	9.9	987.7	33.4	−8.8	960.6	69.4	−17.7
—	10	—	0.9	11.2	3.3	0.9	11.5	3.1	0.9	11.7	2.9	0.8	10.9	3.5	0.8	10.5	3.7
—	25	—	0.8	25.7	1.3	1.0	26.0	0.7	1.2	26.2	0.2	0.7	25.2	1.8	0.6	24.5	2.5
—	50	—	0.1	49.8	2.0	0.3	51.2	0.5	0.6	52.0	−0.8	−0.2	48.0	3.8	−0.3	45.9	5.4
—	100	—	0.3	94.8	2.5	0.8	97.5	−0.5	1.4	99.1	−2.9	−0.0	91.7	5.8	−0.3	87.7	9.1
—	250	—	0.3	253.9	1.1	1.5	260.6	−6.4	3.3	264.5	−13.1	−0.7	245.9	10.1	−1.6	235.3	19.3
—	500	—	0.1	500.0	−0.8	2.9	510.9	−15.9	6.2	518.1	−28.7	−1.9	482.6	16.6	−3.4	462.3	34.5
—	1000	—	−0.7	999.1	−0.9	4.5	1025.8	−31.2	11.4	1041.0	−59.3	−4.7	966.5	33.6	−7.9	925.3	72.0

TABLE 3

Prediction performance of the weight vectors on interpolated numbers of training set

Sample composition[a]	Prediction			Sample composition[b]	Prediction			Sample composition[c]	Prediction		
Cd	Cd	Tl	Pb	Tl	Cd	Tl	Pb	Pb	Cd	Tl	Pb
14.5	11.6	1.4	1.9	14.5	0.9	15.6	2.7	14.5	0.2	1.3	15.5
19	16.1	1.4	1.8	19	0.8	19.9	2.1	19	0.3	1.3	20.2
32.5	28.6	1.4	2.1	32.5	0.6	32.9	1.5	32.5	0.2	1.0	32.7
40	35.1	1.2	2.5	40	0.4	40.2	1.7	40	0.3	0.8	38.5
65	59.1	1.0	3.0	65	0.2	63.3	2.1	65	0.5	0.5	61.0
80	74.4	1.0	3.2	80	0.2	76.8	2.3	80	0.6	0.4	75.7
100	94.7	0.9	3.4	100	0.3	94.8	2.5	100	0.8	0.4	95.2
145	140.9	0.7	3.1	145	0.3	142.5	2.1	145	0.7	−0.0	137.9
190	187.1	0.5	2.8	190	0.3	190.3	1.7	190	0.6	−0.4	180.7
325	321.1	−2.8	1.6	325	0.2	327.7	0.5	325	0.3	−0.6	316.7
400	393.4	−6.0	0.8	400	0.2	401.6	−0.0	400	0.2	−0.2	395.8
650	644.6	−5.7	−0.6	650	−0.1	649.8	−0.8	650	−0.1	0.1	651.5
800	799.4	−1.3	−1.0	800	−0.4	799.5	−0.9	800	−0.3	−0.0	801.8
1000	1005.8	4.5	−1.6	1000	−0.7	999.1	−0.9	1000	−0.1	−0.2	1002.2

[a] $\times 10^{-9}$ M; Tl, Pb absent. [b] $\times 10^{-9}$ M; Cd, Pb absent. [c] $\times 10^{-9}$ M; Cd, Tl absent.

The results of the classification of completely independent real samples of Cd—Pb—Tl mixtures are given in Table 4.

CONCLUSIONS

As can be seen from Table 4 the quantitative predictive ability of the multi-category classifiers for lead, cadmium and thallium is quite good. The determinations are accurate to ±10% over the concentration range 10^{-6}—10^{-8} M, even in cases where the presence of a component cannot be detected visually in the anodic stripping voltammograms (Fig. 1). The method is rather insensitive to shifts of the patterns along the voltage axis such as may be caused by drift of the reference electrode (see Table 2).

The number of four singular values used in the calculations proved to be optimum. With a lower number, all results were poor, whereas the use of higher numbers of singular values decreased the performance in the recognition test, especially when a shift was applied.

TABLE 4

Prediction performance of the weight vectors on separate samples of mixtures of Cd, Pb and Tl

Sample composition (× 10^{-9} M)			Prediction		
Cd	Tl	Pb	Cd	Tl	Pb
25	25	25	22.2	25.1	22.9
25	25	25	22.5	25.2	23.7
1000	1000	1000	1044.2	1083.6	1044.7
50	1000	50	48.1	1051.8	65.4
50	1000	50	48.4	1051.4	59.1
100			92.4	0.1	1.7
100			90.8	—0.5	1.7
			0.4	1.8	2.3
			0.6	2.4	2.6
			0.6	1.6	3.0
		10	0.4	1.2	11.3
		100	0.7	0.6	93.7
	100		1.1	99.0	—2.9
	W50 C		0.3	0.1	0.2

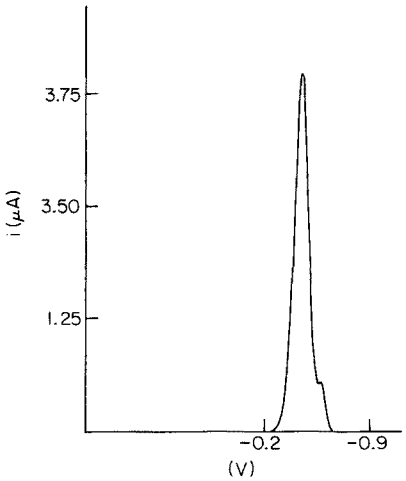


Fig. 1. A.s.v. of 10^{-6} M Tl, 5×10^{-8} M Cd and 5×10^{-8} M Pb.

The authors are indebted to J. Nieukoop for preparing the manuscript and to Prof. E. A. M. F. Dahmen for his interest.

REFERENCES

1 B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 1945.
2 S. R. Lowry, T. L. Isenhour, J. B. Justice, F. W. McLafferty, H. E. Dayringer and R. Venkataraghavan, Anal. Chem., 49 (1977) 1720.
3 C. L. Wilkins and T. R. Brunner, Anal. Chem., 49 (1977) 2136.
4 Q. V. Thomas, R. A. De Palma and S. P. Perone, Anal. Chem., 49 (1977) 1376.
5 P. C. Jurs, B. R. Kowalski and T. L. Isenhour, Anal. Chem., 41 (1969) 21.
6 B. R. Kowalski, P. C. Jurs, T. L. Isenhour and C. N. Reilley, Anal. Chem., 41 (1969) 695.
7 D. D. Tunnicliff and P. A. Wadsworth, Anal. Chem., 45 (1973) 12.
8 Y. C. Ho and R. L. Kashyap, IEEE Trans. Electron. Comput., EC-14 no. 5, (1965) 683.
9 A. Albert, Regression and the Moore—Penrose Pseudo-inverse, Academic Press, New York, 1972.
10 A. Ben Israel and T. N. E. Greville, Generalized Inverses; Theory and Applications, Wiley, New York, 1974.
11 P. A. Businger and G. H. Golub, Comm. ACM 12 (1969) 564.
12 G. H. Golub and W. Kahan, SIAM J. Numer. Anal., 2 (1965) 205.
13 S. Söderström, SIAM J. Numer. Anal., 11 (1974) 61.
14 J. H. Wilkinson and C. Reinsch, Handbook for Automatic Computation, Vol. II, Linear Algebra, Springer Verlag, Berlin, 1971, p. 134—151.
15 T. Y. Young and T. W. Calvert, Classification, Estimation and Pattern Recognition, Elsevier, New York, 1973.
16 J. Sinko and J. Doležal, J. Electroanal. Chem., 25 (1970) 53.
17 G. H. Golub and C. Reinsch, Numer. Math., 14 (1970) 403.

Analytica Chimica Acta, 103 (1978) 167—173
Computer Techniques and Optimization

QUANTITATIVE ANALYSIS FOR POLYCYCLIC AROMATIC HYDROCARBONS BY SPECTRAL DECOMPOSITION OF MOLECULAR FLUORESCENCE

HARVEY S. GOLD, CARL E. RECHSTEINER, JR. and RICHARD P. BUCK*
William R. Kenan Jr. Laboratories of Chemistry, University of North Carolina, Chapel Hill, NC 27514 (U.S.A.)
(Received 14th March 1978)

SUMMARY

Analysis of the molecular fluorescence spectra of several polycyclic aromatic hydrocarbons (PAH) by the technique of spectral decomposition is discussed. Peak characterization parameters obtained by this technique are utilized to obtain quantitative measurement of a mixture containing benzidine and dibenzochrysene in solution. Application of the technique to a large number of reference compounds is discussed as an aid to routine identification of suspected carcinogens and other compounds containing fluorophores.

The ever-increasing number of polycyclic aromatic hydrocarbons (PAH) that have been linked with carcinogenesis has caused widespread attention to be focused on their analysis. As early as 1933, Cook et al. [1, 2] utilized fluorescence to identify a carcinogenic PAH substituent of coal tar, after earlier investigations of Kennaway and Heiger [3, 4] which noted the fluorescence of coal tar. By 1953, over 200 PAH were known to be carcinogenic. The high sensitivity of fluorimetric methods has resulted in their continued application to PAH analyses [5]. To date, fluorescence measured in liquid solution has been largely unsatisfactory for quantitative analysis of mixtures of fluorophores. Like u.v.—visible absorption processes, the fluorescence of aromatic molecules exhibits generally broad, featureless spectra, thus traditionally leading to severe problems of spectral overlap in the case of mixtures. A second problem is the effect of quenching and intermolecular energy transfer which serves to distort the relationship between observed fluorescence intensity and species concentration.

There have been reports of successful analyses for carefully selected conditions [6] and systems [7, 8]. Recent work [9] has resulted in a method of fluorescence analysis which relies on spectral decomposition, a curve-fitting technique that has been widely applied to a range of absorption spectral types such as infrared spectroscopy [10—13] and recently to some forms

of emission spectroscopy [14]. This method uses an iterative least-squares procedure to find and characterize component bands of the fluorescence envelope. Peak descriptors consisting of peak location, peak-width parameters, and intensities are obtained from SPECSOLV, a FORTRAN computer program [9, 14, 15]. These descriptors allow unambiguous identification of mixture components; this in turn permits quantitative results to be obtained from the decomposed spectrum.

EXPERIMENTAL

Instrumentation

A Hitachi MPF-2A fluorescence spectrometer was used in the direct mode to obtain the emission spectra of the sample solutions; 10-nm emission and excitation slits were used. Digital data were recorded at 5-nm intervals from a Keithley model 160 digital multimeter which was used in place of the standard chart recorder.

Reagents

Spectroquality cyclohexane (MC/B), absolute ethanol, and A.C.S. reagent-grade methanol were used as solvents. Benzidine (Baker Analyzed), heptaphene (Aldrich), and dibenzochrysene (Aldrich) were quantitatively dissolved in the appropriate solvent with ultrasonic shaking.

Data analysis

Digitized spectrofluorimetric data were analyzed by the SPECSOLV program [9, 15].

RESULTS

The fluorescence spectra of benzidine, heptaphene, and dibenzochrysene were corrected for background and successfully decomposed by SPECSOLV. The spectra of benzidine (Fig. 1a) and dibenzochrysene (Fig. 1b) each consisted of only one peak, whereas the spectrum of heptaphene (Fig. 1c) is composed of three component peaks. As a check on the number of component peaks present, decomposition into 2—4 peaks was tried for benzidine and dibenzochrysene, while 2 and 4—6 peaks were tested for heptaphene. In all cases, these attempts resulted in obviously spurious peaks and, significantly, failure of the program to converge when the iteration limit was reached. This provides an internal check against over-specification or under-specification of the number of peaks to be used in the decomposition process. The peak parameters which describe the spectra presented as Fig. 1 are summarized in Table 1.

As initial models, the three compounds used clearly illustrate that even "featureless" spectra can be reported by descriptors of high information content. A cursory survey of a large number of fluorescence spectra of aromatic

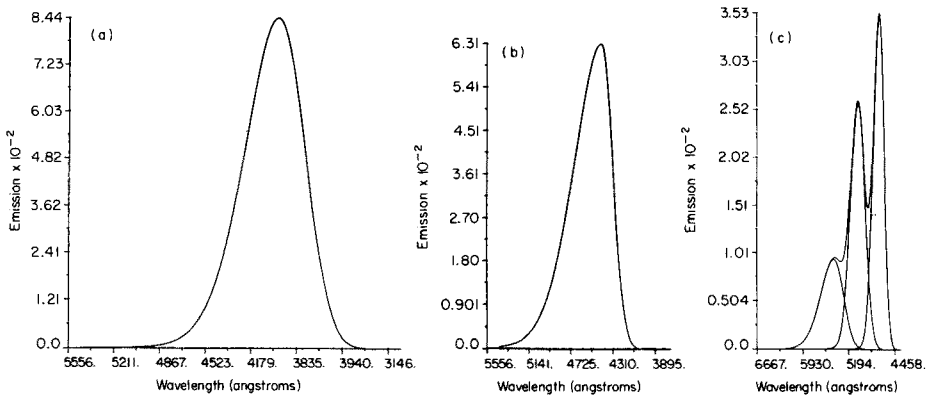


Fig. 1. Decomposed fluorescence spectrum of benzidine (a), dibenzochrysene (b) and heptaphene (c). Calculated spectra with bi-Gaussian components.

TABLE 1

Peak parameters of reference compounds

	Position (nm)	Half-widths (nm)		Intensity
		(−)	(+)	
Benzidine	393.1	28.13	25.13	8.44
Dibenzochrysene	442.6	29.31	14.02	6.31
Heptaphene	471.7	9.98	9.93	3.50
	505.0	12.86	13.93	2.57
	542.6	23.76	18.49	0.94

molecules [16] shows that most are at least qualitatively similar to the range of species analyzed to date by SPECSOLV in this and a previous study [9]. In the latter, the ability to distinguish and identify solvent peaks and to assign peaks in a mixture to a particular species was demonstrated, and concentration effects on peak characterization parameters were investigated.

The current study extends the analysis to two species, benzidine and dibenzochrysene, the spectra of which each exhibit only one asymmetric peak. Since spectral decomposition has evolved as a method of resolving overlapping peaks, its application to a single peak spectrum may well be questioned in terms of utility and necessity. However, the technique does yield four spectral descriptors which totally characterize the fluorescence spectrum of the single fluorescence peak compounds much as $4n$ (where n is the number of peaks in the spectrum) descriptors completely characterize a spectrum with overlapping peaks. While the peak location in a single-peak spectrum may be obvious without decomposition, such factors and par-

ameters as asymmetry and precise numerical measures of half-widths are not. Of significance is the demonstration that SPECSOLV did not converge on equivalent solutions which involve introduction of more peaks, when the fluorescence spectrum in fact consisted of only one peak. This problem of ambiguous solutions has been a major problem with some past decomposition algorithms, and has been discussed extensively [17—19].

Spectral characterization can be utilized to great advantage for both qualitative and quantitative analysis. The peak descriptors are readily used to identify fluorescing materials, and provide an ideal data base for computer analysis. Spectral decomposition, by providing a relatively small number of peak descriptors per compound, is an efficient and easily implemented method for computer identification of fluorescent species. The feasibility of this has been demonstrated by Miller and Faulkner [20] in a study which developed a data base with a first derivative peak-searching algorithm.

To be generally applicable, all reference spectra should be run on a corrected spectrofluorimeter, because this effectively eliminates instrument dependence. This capability is not currently available in this laboratory; nevertheless, the instrument-dependent data base developed to date is sufficient to demonstrate feasibility.

A major concern in analytical chemistry is the determination of materials in mixtures. This is commonly accomplished after a separation stage. However, spectral decomposition not only provides information on species identity, but also yields peak-intensity data which can be transformed to yield highly accurate determinations of all absorbing or fluorescing species present in the mixture. Thus the requirements regarding separation of all materials are reduced. While this has always been true when no material interfered with the determination of any other, accurate determination of spectroscopically interfering substances without prior separation was elusive at best. Spectral decomposition offers a way round this problem.

The determination of benzidine—dibenzochrysene mixtures is an illustrative example. The fluorescence spectrum of this mixture exhibits severe overlap. The digitized spectrum of a particular concentration of these species is shown in Fig. 2(a). The 5×10^{-2} mg ml^{-1} solution of dibenzochrysene had a peak intensity of 631, while benzidine at 7×10^{-2} mg ml^{-1} had an intensity of 844 (Table 1). It is also possible to define height—width parameters (HWP) of 136.7 and 224.8 for dibenzochrysene and benzidine, respectively, where HWP is given by the expression

$$\text{HWP} = \{[\text{half-width}(+) + \text{half-width}(-)]/2\} \times \text{peak intensity}$$

The result of decomposition of a binary mixture of benzidine and dibenzochrysene is shown in Fig. 2(b); the actual intensity of the peak at 393.3 nm is significantly diminished, as a result of being located on the shoulder of the peak centered at 442.3 nm. The decomposition process resulted in peaks with the intensities and HWP values given in Table 2. The peak at 442.3 nm is unambiguously assignable to dibenzochrysene, while that at 393.3 nm is

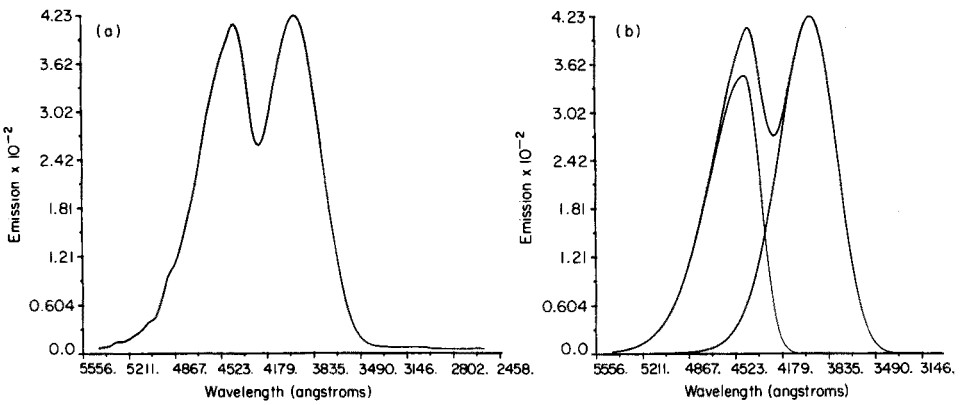


Fig. 2. Fluorescence spectrum of a benzidine—dibenzochrysene mixture (a) and of the same spectrum after decomposition (b).

TABLE 2

Component values for benzidine—dibenzochrysene mixture

Peak	Position (nm)	Half-widths (nm)		Intensity	HWP	Species
		(−)	(+)			
1	393.3	25.55	25.43	4.22	107.6	Benz.
2	442.3	30.25	14.34	3.48	77.5	Dibenz.

due to benzidine. By comparison of the peak intensities in the binary case to both the intensities and known concentrations of the single-component references (Table 1), the concentrations of dibenzochrysene and benzidine in the test mixture are 2.76×10^{-2} mg ml^{-1} and 3.50×10^{-2} mg ml^{-1}, respectively. The solution actually contained 2.75×10^{-2} mg ml^{-1} of dibenzochrysene and 3.50×10^{-2} mg ml^{-1} of benzidine. Use of the HWP values gives results of 2.83×10^{-2} mg ml^{-1} and 3.35×10^{-2} mg ml^{-1}; the latter value for benzidine is low, probably because of the effects of noise on the course of decomposition. Neither the peak height nor the HWP method takes into account the effects of possible intermolecular energy transfers; as all spectra were run on the same instrument, artifacts caused by non-uniform source and detector responses effectively cancel. The agreement between known and experimentally determined values is quite good.

A number of benzidine—dibenzochrysene mixtures were analyzed to evaluate the range of this method for any mixture of A and B. There should in principle be two logical limits. These constraints are: 1) the relative ratio of A to B or of B to A is sufficiently low that the peak used to determine A or B approaches the signal level attributable to residual baseline noise;

and 2) the concentration of A and/or B is sufficiently high that concentration quenching or other non-linear effects occur. The latter is, of course, a matter of general concern in fluorescence spectrometry and is not peculiar to decomposition analysis. The former case is quite significant, for it serves to provide a rough dynamic range for quantitative analysis by this method. To address this issue, various concentrations of dibenzochrysene were studied in the presence of a uniform benzidine concentration (2×10^{-1} mg ml^{-1}). Analysis of the spectrum (Fig. 2a) showed that the residual noise was 5.00×10^{-4} emission units, approximately 1.2% of the maximum peak intensity of benzidine, corresponding to 2.3×10^{-4} mg ml^{-1} of this compound. When the concentration of benzidine was increased by a factor of 10, the residual baseline noise decreased by roughly a factor of 10, to about 0.1%. The quantity of benzidine represented by this noise was essentially unchanged.

The initial quantitative results indicate that a dibenzochrysene concentration of 2.43×10^{-1} mg ml^{-1} would be required to give a dibenzochrysene peak equal in intensity to the 2.0×10^{-1} mg ml^{-1} benzidine peak. Assuming a minimum acceptable signal-to-noise ratio of 2, a peak of 0.2% of this full scale value could ideally be detected. This would correspond to 4.8×10^{-4} mg ml^{-1} of dibenzochrysene. In practice, this was not achieved. Solutions ranging from 2.43×10^{-1} mg ml^{-1} to 2.43×10^{-3} mg ml^{-1} were prepared and analyzed. The successful quantitative results are shown in Table 3; reliable results were obtained when the emission intensity was at least 5% of the maximum emission. This then represents a dynamic concentration range of 40:1 for each component, since either could have represented the full-scale peak. The intensity of the benzidine peak was constant within 1%, consistent with the fact that the actual solution concentration was held fixed at 2.0×10^{-1} mg ml^{-1}. When dibenzochrysene concentrations that gave peaks with intensities of less than 123 emission units were analyzed, decomposition results were erratic (errors in peak positions or intensities were common and unpredictable).

While the benzidine—dibenzochrysene system provides a convenient example of a mixture with only two peaks, the results are more broadly applicable. Similar studies were performed on a mixture of benzidine and

TABLE 3

Analysis for dibenzochrysene in the presence of benzidine (0.2 mg ml^{-1})

Concn. taken ($\times 10^{-2}$ mg ml^{-1})	Peak intensity at 442.3 nm (arbitrary units)	Concn. calcd. ($\times 10^{-2}$ mg ml^{-1})	Deviation (%)
24.3	2557	24.3	0.0
18.1	1885	17.9	−1.1
12.1	1284	12.2	+0.8
2.43	256	2.43	0.0
1.81	193	1.83	+1.0
1.18	123	1.17	−0.5

heptaphene and on a mixture of dibenzochrysene and heptaphene. Both of these are systems with four peaks, the latter exhibiting severe overlap. In addition, various binary combinations of these three compounds and four previously characterized materials [9] (anthracene, naphthalene, naphthacene, and rubrene) were analyzed to consider cases with large numbers of overlapping peaks. Generally the requirement observed for the benzidine—dibenzochrysene case is followed: for accurate quantitative results, the emission intensity of the most intense peak of the minor (based on emission intensity, not on concentration) constituent must be at least 5% of the intensity of the most intense peak present in the spectrum.

This work was presented in part at the North Carolina Section of the American Chemical Society Meeting-in-Miniature, Chapel Hill, NC, April, 1977. Support by the National Science Foundation under grant MPS75-00970 is gratefully acknowledged.

REFERENCES

1 J. W. Cook, C. L. Hewett, and I. Heiger, J. Chem. Soc., (1933) 395.
2 J. W. Cook and C. L. Hewett, J. Chem. Soc., (1933) 398.
3 E. L. Kennaway, Brit. Med. J., 3622 (1925) 1.
4 E. L. Kennaway and I. Heiger, Brit. Med. J., 3622 (1930) 1044.
5 E. Sawicki, Talanta, 16 (1969) 1231.
6 T. L. Pasby, in A. J. Pesce (Ed.), Fluorescence Spectrometry, Dekker, New York, 1971, 149—201.
7 P. Einarsson, H. Hallman, and G. Jonsson, Med. Biol., 53 (1975) 1.
8 O. Lindvall, A. Bjorklund, and B. Talch, J. Histochem. Cytochem., 23 (1975) 697.
9 C. E. Rechsteiner, H. S. Gold, and R. P. Buck, Anal. Chim. Acta, 95 (1977) 51.
10 R. D. Fraser and E. Suzucki, Anal. Chem., 41 (1969) 37.
11 K. S. Sheshradi and R. N. Jones, Spectrochim. Acta, 19 (1963) 1013.
12 J. Pitha and R. N. Jones, Can. J. Chem., 44 (1966) 3031.
13 R. P. Young and R. N. Jones, Chem. Rev., 71 (1971) 219.
14 H. S. Gold, C. E. Rechsteiner, and R. P. Buck, Anal. Chem., 48 (1976) 1540.
15 H. S. Gold, SPECSOLV — A Generalized Spectral Decomposition Program, Library Service Series Document No. LS-301, Research Triangle Park (N.C.), Triangle Universities Computation Center, 1976.
16 I. B. Berlman, Handbook of Fluorescence Spectra of Molecules, Academic Press, New York, 1971.
17 J. R. Beacham and K. L. Andrew, Opt. Soc. Am. J., 61 (1971) 231.
18 J. W. Perram, J. Chem. Phys., 49 (1968) 4225.
19 A. R. Davis, D. E. Irish, R. B. Roden, and A. J. Weerheim, Appl. Spectrosc., 26 (1972) 384.
20 T. C. Miller and L. R. Faulkner, Anal. Chem., 48 (1976) 2083.

Analytica Chimica Acta, 103 (1978) 175—187
Computer Techniques and Optimization

SIMULATION OF ELECTRON SPIN RESONANCE SPECTRA BY FAST FOURIER TRANSFORM

A Novel Method of Calculating Spectra to Include Isotopic Substitution, Superhyperfine Coupling, Instrument Time Constant and Modulation Broadening in Fluid and Polycrystalline Media

J. C. EVANS*, P. H. MORGAN and R. H. RENAUD
Chemistry Department, University College, Cardiff CF1 1XL (Great Britain)
(Received 3rd January 1978)

SUMMARY

A computer program is described for the rapid calculation of solution and polycrystalline electron spin resonance spectra of systems containing one unpaired electron. The calculation time is virtually independent of the number of e.s.r. transitions considered, e.g., morphamquat radical cation (3025 lines) requires 4.5 s. Second-order corrections and line-width anisotropy can be included in the simulation. The graphical output may be matched to the output of any e.s.r. spectrometer. Instrumental parameters, e.g. modulation amplitude and time constant, are accounted for in the calculation of the simulated spectrum which enables exact comparison between experimental and simulated spectra. To accommodate mixtures of paramagnetic species, a spectrum addition facility is provided; the output may be presented as an absorption or as any derivative. Spectra originating from isotopically substituted molecules may be calculated routinely and quickly, without the necessity for prior calculation of the relative contributions of the various combinations of isotopic nuclei.

Several examples are given, illustrating the usefulness of this program in extracting spectral information under most experimental conditions.

The recording of a spectrum is only the first step towards the final aim of the evaluation of spectroscopic constants. A tedious and often complicated analysis must be done in order to extract this information from recorded spectra. Generally speaking, it is a problem of reducing the amount of data to a few relevant constants. The electron spin resonance (e.s.r.) spectra of organic free radicals resulting from the isotropic or anisotropic interaction of the unpaired electron with the magnetic fields may be represented in terms of a superposition of elementary functions (or lines) whose shape is known approximately [1]. The relative weights and positions of the lines may be specified by a set of splitting constants representing a set of convolution operations on the elementary line-shape function. The determination of the splitting constants from experimental spectra may be easily accomplished provided that the number of lines is small and there is no appreciable overlap.

The interpretation of more complicated spectra can, however, be a difficult task. The modern high-speed digital computer provides a valuable tool for the interpretation of such spectra. Chen et al. [2] have devised a program for the determination of the ratio of the nitrogen splitting constants in *p*-substituted 1,1-diphenyl-2-picrylhydrazyl. Spectral simulation programs which reconstruct a spectrum from a trial set of splitting constants have been prepared by Stone and Maki [3] for isotropic e.s.r. spectra and by Lefebvre and Maruani [4] for e.s.r. spectra of amorphous solid samples. Several methods of least-squares approximation which are applicable to the interpretation of many types of spectra by a digital computer have been presented [5—7]. A digital computer program for least-squares analysis of multi-line spectra by using tentative splitting constants has been prepared [8].

This paper reports a novel method for the fast simulation of e.s.r. spectra of organic radicals in solution, transition metal complexes ($S = \frac{1}{2}$) both in solution and in the polycrystalline state for molecules of monoclinic or higher symmetry, by means of the fast Fourier transform method. Where natural abundances of isotopes give rise to mixed spectra, e.g. ^{13}C or ^{29}Si, this method removes the need to calculate abundances of various isotope combinations and includes all possible isotopic contributions with very little loss of speed. Mixed spectra from various paramagnetic species, other than those resulting from isotropic substitution can be simulated by the addition of the relevant percentage amounts of the individual spectra. If the necessary parameters can be estimated, anisotropic line-width dependence on M_I (the nuclear spin quantum number) caused by incomplete averaging can be accommodated. Second-order terms in hyperfine interactions can be allowed for, in calculations of line positions.

Applications of this method to the analysis of spectra obtained from (a) a vanadium(IV)—aluminium complex [9], (b) the radical cation of paraquat [10], and (c) the radical anion of bistrimethylsilyldiacetylene [11] with potassium as counter ion, which contains ^{13}C and ^{29}Si isotopic splitting, illustrate the various features of the program and show the usefulness of fast and accurate analysis of spectra.

A Varian E3 e.s.r. spectrometer with 100 kHz modulation was used for the measurements and recording of the spectra. An ICL System 4-70 computer with a main store of 768 kbytes running on a multi-access system, MULTIJOB, was used to construct the simulations.

THEORY

Previous workers [12] have described simulation methods which calculate line intensity and position information prior to superimposing some specified line-shape function. These programs are designed to produce ideal spectra, and as the instrumental distortions produced by modulation and time constant effects change line positions and intensities, it is preferable

to include them in the simulation. The programs described by Negoita et al. [12] use path A (see Fig. 1) to obtain the simulated spectrum, which does not include modulation and instrumental conditions, and this effectively involves discretely evaluating the following integral [13]:

$$A = \int_{-\infty}^{+\infty} s(\tau) l(t-\tau) \mathrm{d}\tau \tag{1}$$

where τ is a dummy variable, t is the time domain variable, l is the line-shape function, and s denotes a sum of delta functions representing the "stick" spectrum. For this purpose, a separate simulation is made for each value of t considered, whereas the Fourier transform (FT) procedure (path B) involves evaluation of integral B

$$B = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} s(t) \exp(-2\pi i f t)\, \mathrm{d}t \int_{-\infty}^{+\infty} l(t) \exp(-2\pi i f t)\, \mathrm{d}t \right] \exp(2\pi i f t)\, \mathrm{d}f \tag{2}$$

where f is the frequency domain variable.

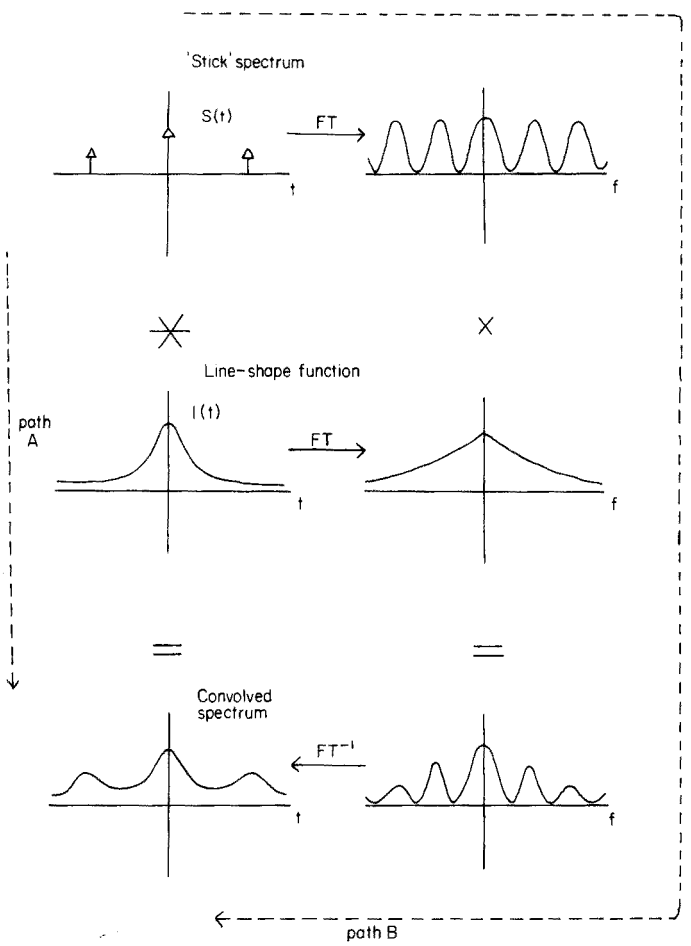


Fig. 1. An illustration of the use of the Fourier transform convolution theorem. * Denotes the operation of convolution; FT and FT^{-1} denote forward and inverse Fourier transformation.

There are several features of the Fourier transform that enable simulation to be carried out particularly efficiently. First, probably the most time-consuming process in simulation, especially when a large number of transitions is being considered, is the application of the line shape. The Fourier transform counterpart of this convolution operation is the multiplication of the Fourier transforms of the stick spectrum and the line-shape function. As the latter is analytically known for both Lorentzian and Gaussian curves, either can be used with equal ease. The result is the application of the line-shape function to all lines in one operation. Secondly, in the case of first-order spectra, such as those of many organic radicals, the Fourier transform of the stick spectrum for a nucleus of any nuclear spin is also known, and a consequence of the convolution theorem is that, if n equivalent nuclei are present, the appropriate function need only be raised to the n-th power to generate the correct line intensities (see Appendix). Thirdly, the effects of time constant and modulation are also convenient analytical functions in the Fourier transform domain. Fourthly, the Fourier transform of an n-th derivative is obtained by multiplying the Fourier transform of the absorption spectrum by $(2\pi if)^n$ where $i = (-1)^{1/2}$, and f is the Fourier transform counterpart of the variable t (in this case, the external magnetic field) in real space. Thus the output may be presented as the first derivative or second derivative, etc., with equal ease.

Further, the Fourier transform approach provides a very simple method of dealing with isotopic mixtures. Consider, for example, the radical anion $[(CH_3)_3Si{-}C{\equiv}C{-}C{\equiv}C{-}Si(CH_3)_3]^{\cdot-}K^+$ obtained by passing a THF solution of the diacetylene over a potassium film [11]. The e.s.r. spectrum of this species shows splitting from ^{29}Si and ^{13}C; when normal methods are used, it is necessary to pre-calculate the various probabilities of any molecule containing one or more of any of the known isotopes ^{12}C, ^{13}C, ^{28}Si, ^{29}Si, then simulate each contribution, and add them. The Fourier transform of the above spectrum is however readily calculated as:

$$[W(^{12}C)FT(^{12}C) + W(^{13}C)FT(^{13}C)]^6 \times [W(^{28}Si)FT(^{28}Si) + W(^{29}Si)FT(^{29}Si)]^2 \times [FT(^1H)]^{18} \times [FT(^{39}K)] \times FT \quad (3)$$

where FT(nucleus) is the Fourier transform function of the splitting from a particular nucleus. If the nucleus has no nuclear spin, then FT(nucleus) = 1. The exponents refer to the number of geometrically equivalent nuclei. W(nucleus) is the relative abundance of the particular nucleus under consideration, e.g. 0.9889 for ^{12}C, 0.0111 for ^{13}C, etc.

It is easier to see how this works by considering just one set of isotopes, ^{28}Si, ^{29}Si, and expanding the appropriate part of expression (3):

$$W(^{28}Si)^2FT(^{28}Si)^2 + 2W(^{28}Si)W(^{29}Si)FT(^{28}Si)FT(^{29}Si) + W(^{29}Si)^2FT(^{29}Si)^2$$

$W(^{28}Si)^2$ is the probability of both silicon atoms being ^{28}Si, whereas $2W(^{28}Si)W(^{29}Si)$ is the probability of one silicon atom being ^{29}Si. When this argument

is extended to the multiple isotope mixture, it can be seen that in one operation, the Fourier transform permits effective consideration of all possible contributions, even that arising from molecules containing two ^{29}Si and six ^{13}C nuclei.

The Fourier transform always produces spectra of equal area, hence there is no need to calculate multiple integrals of the spectra of components in mixtures; one simply adds the estimated percentages of the spectra together. Thus (a) spectra of any derivative may be added together with equal ease; and (b) since no numerical integration is required, the amounts of each component present can be determined to high accuracy, by comparison of the calculated and experimental spectra.

A final advantage is that particularly efficient algorithms are already available for the calculation of the Fourier transform and its inverse [14, 15].

Together, these advantages lead to a very fast method for simulation of e.s.r. spectra. A fast Fourier transform simulation program was written; this included the various features described above. It was found that all solution spectra could be calculated in 6 s or less from 2048 data points on an ICL 4-70 computer. A subroutine for calculating polycrystalline spectra was included. A and g tensors were assumed to have only diagonal entries, with the exception of A_{xy}, A_{yx} specified by a rotation of the principal axis of the A tensor in the x—y plane. This corresponds to a minimum symmetry requirement of C_2 in the system being simulated. Superhyperfine interaction arising from one set of magnetically equivalent nuclei may be allowed for, provided that the presence of the ligand(s) does not reduce the symmetry to below C_2.

APPLICATIONS TO ELECTRON SPIN RESONANCE SPECTRA

The program was used to analyse e.s.r. spectra from transition metals in solution and in the polycrystalline state and also complicated organic radical cation and anion spectra. In all the above cases, isotopic effects could be incorporated routinely.

Figure 2(a) shows a simulated e.s.r. absorption spectrum based on parameters for dichloro-bis(η-cyclopentadienyl)vanadium(IV). The eight line positions (I = 7/2) are calculated correct to second order and show line-width dependence on the nuclear spin quantum number. Any derivative spectrum can be obtained easily, and Fig. 2(b) shows the second derivative spectrum of Fig. 2(a). The line shape of the second derivative (Fig. 2b) is considerably distorted because of the simulated time constant which, as is seen, has very little effect on the absorption spectrum Fig. 2(a). This difference is due to the fact that the time-constant circuit affects predominantly the high-frequency information in the signal. Obtaining the derivative is equivalent to high-pass filtering, which exaggerates the high-frequency distortion of the time-constant circuit. Therefore, in order to obtain sharper lines for greater accuracy of the hyperfine-splitting constants and g values, short-time constants should be used in the derivative spectra.

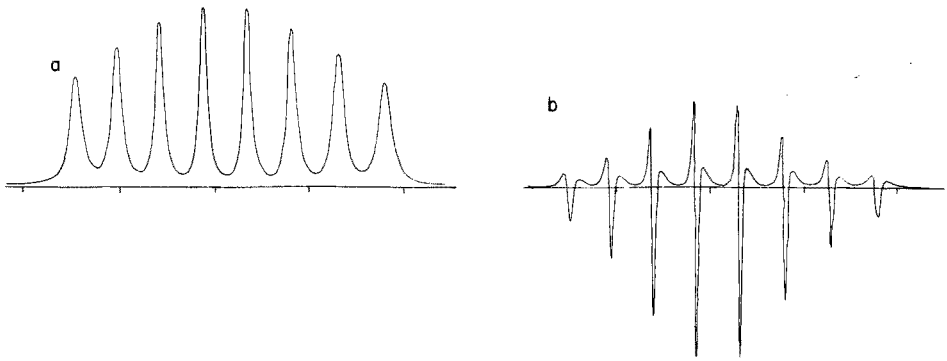


Fig. 2. (a) Computed absorption spectrum for a solution of Cp_2VCl_2 at room temperature from parameters as in ref. [9]. (b) Second-derivative presentation of Fig. 2(a).

The e.s.r. spectrum obtained by passing a solution of dichloro-bis(η-cyclopentadienyl)vanadium(IV) over an aluminium chloride film, is shown in Fig. 3(a) and its interpretation has already been published [9]. This spectrum shows eight hyperfine lines from the vanadium nucleus, A_{iso}^{V} = 7.4 mT, each line being split into a further 6 lines by one aluminium nucleus (A_{iso}^{Al} = 0.63 mT).

The use of this program has confirmed the analysis made previously [9] by simulating the e.s.r. spectrum (Fig. 3b). Therefore superhyperfine coupling can be easily accommodated, together with line-width dependence on M_I and second-order effects. It is interesting to note how the simulated spectrum faithfully reproduces the progressive deterioration of the aluminium superhyperfine coupling on the high field lines.

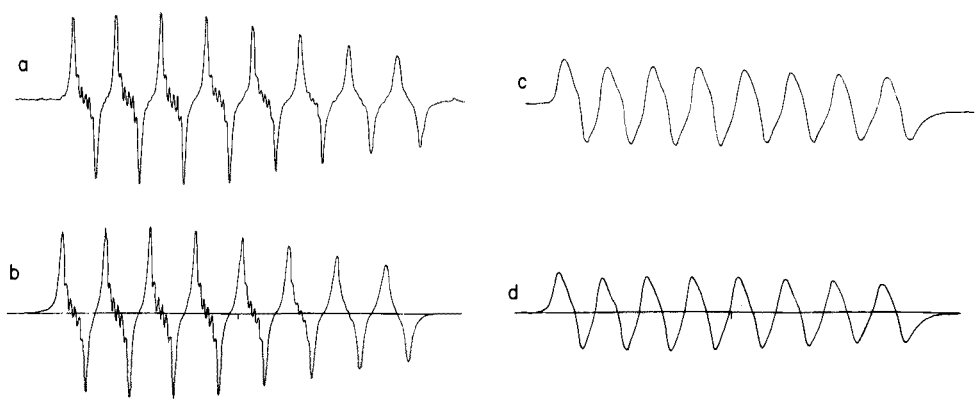


Fig. 3. (a) Experimental spectrum of vanadium(IV) complex in solution at room temperature, showing superhyperfine coupling from one Al nucleus [9]. (b) Computed spectrum of (a) with parameters from ref. [9]. (c) Experimental spectrum of (a) with time constant and modulation amplitude increased from 0.3 to 3 s and from 0.05 to 1.0 mT respectively. (d) Computed spectrum of (c) taking into account the increased time constant and modulation amplitude.

Figures 3(c) and (d) show the experimental and computed spectra for the sample shown in Fig. 3(a). The time constant and modulation amplitude were increased from 0.3 to 3 s and 0.05 to 1.0 mT, respectively. A very good fit is obtained, showing that instrumental factors are easily incorporated, a provision not available as far as is known in other simulation programs.

In order to demonstrate that the program can simulate polycrystalline e.s.r. spectra, a series of such spectra is shown in Fig. 4.

The program can also be used effectively to simulate polycrystalline spectra that arise in an experimental situation. Figure 5 shows a simulation of the spectrum obtained from a 0.01 M solution of dichloro-bis(η-cyclopentadienyl)-vanadium(IV) in chloroform, as well as the experimental spectrum. The A

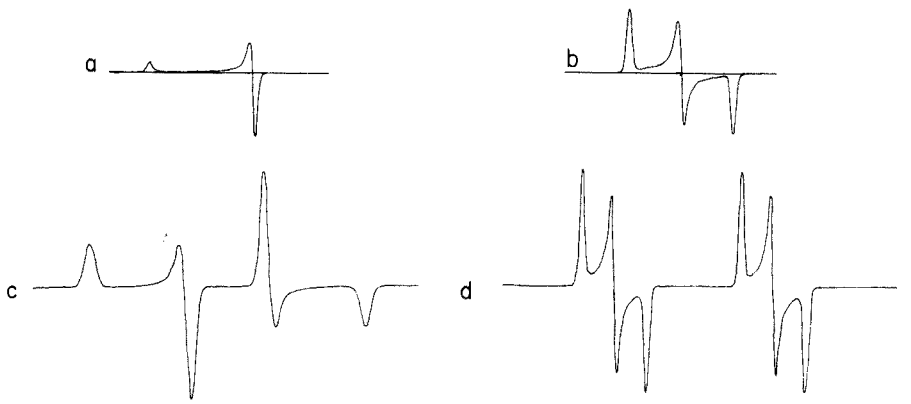


Fig. 4. Hypothetical polycrystalline e.s.r. spectra. (a) $S = \frac{1}{2}$; axial symmetry, no hyperfine coupling ($I = 0$); (b) $S = \frac{1}{2}$; orthorhombic symmetry; no hyperfine coupling; (c) $S = \frac{1}{2}$, $I = \frac{1}{2}$, $g_z \neq g_y = g_x$; $A_z = A_y = A_x$ (isotropic A); (d) $S = \frac{1}{2}$, $I = \frac{1}{2}$, $g_z = g_y = g_x$; $A_z \neq A_y \neq A_x$ (anisotropic A).

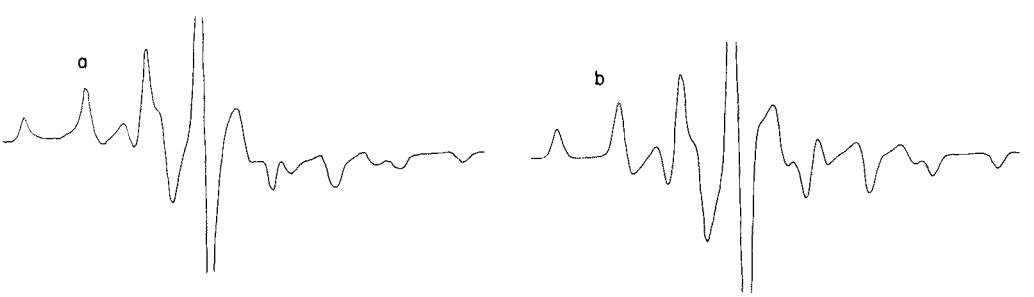


Fig. 5. (a) Experimental polycrystalline e.s.r. spectrum from a 0.01 M solution of dichloro-bis(η-cyclopentadienyl)vanadium(IV) in chloroform. (b) Simulation of this spectrum for the following parameters: line-width, 2.0 mT; $g_z = 2.0013$, $g_x = 1.9802$, $g_y = 1.9695$; $A_z = 2.06$ mT, $A_x = 8.04$ mT, $A_y = 12.60$ mT.

and g tensors were assumed to have coincident principal axes. As can be seen the agreement between the simulated and the experimental spectra is excellent.

The lowest symmetry for which the program is usable is monoclinic symmetry. The big advantage of this program over others published is its speed of computation. Considerable saving in time is achieved in this case by first calculating and adding the "stick" spectra for all orientations of the crystallite, and then superimposing the line shape by means of the Fourier transform, because pointwise multiplication in the transform domain is inherently faster than convolution. The e.s.r. spectra of radical cations of reduced substituted 4,4′-bipyridilium salts have been extensively studied [10, 16].

Figure 6(a) shows an e.s.r. spectrum of the radical cation derived from the 1e-reduction of the 1,1′-dimethyl-4,4′-bipyridilium dichloride salt (paraquat). The e.s.r. spectrum was obtained under optimum conditions which gave good resolution (about 150 lines observed out of a total of 875 lines), therefore the various proton and nitrogen coupling constants can be obtained. Figure 6(b) is a computed spectrum based on the parameters obtained from Fig. 6(a) [10]; this shows a very good fit thus confirming the analysis. The only significant improvement would be to include isotope coupling (^{13}C) in the simulation. The advantage of using this program over others is that the computation time for first-order spectra is virtually independent of the number of spectral lines, e.g. the spectrum shown in Fig. 6(b) takes 9 s using a conventional convolution routine and 4 s by the present Fourier transform technique. The difference shows itself more readily for even more complex spectra, e.g. [morphamquat]$^{+\cdot}$ takes 25 s compared to 4.5 s with the present program [16].

In this case the effects of time constant and modulation can be particularly pronounced. This arises from the large number of transitions normally involved and the relatively small line-width (typically about 0.01 mT). Here again, the instrumental effects act as low-pass filters and so remove high-frequency information, thereby distorting the spectrum, The inclusion of these parameters in the simulation enables a good fit to be obtained with the experimental spectrum even under adverse conditions, where high modulation and time constant are required.

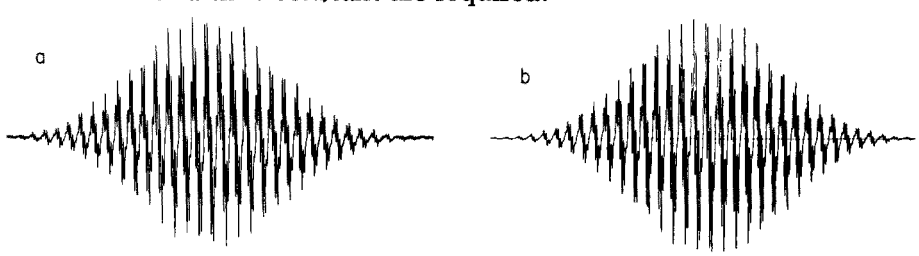


Fig. 6(a) E.s.r. spectrum of the paraquat radical cation in methanol at room temperature; (b) computed spectrum of paraquat radical cation with line-width 0.0155 mT, T_c 0.3 s, modulation amplitude 0.005 mT, time scan 1 h, and with the hyperfine coupling constants in ref. [10].

Figure 7 illustrates the deterioration in the resolution of the experimental spectrum as the ratio of the time constant to scan time is increased. It can be seen that the simulation follows accurately the change in the spectrum, simply by modifying the value of the time constant as obtained from the machine (compare spectra a—f).

The need to consider isotopic splitting arose from an e.s.r. study of the radical anion formed by the reduction of a tetrahydrofuran solution of bis(trimethylsilyl)diacetylene with potassium [11]. The e.s.r. spectrum of the radical anion appeared to show a central group of lines, from the eighteen equivalent protons (the molecule contains two ^{28}Si nuclei), with a similar less intense band on either side (Fig. 8a). These were first interpreted as further interaction with a ^{29}Si nucleus (natural abundance, 4.7%). However, simulation and close measurement showed that, in fact, neither of these assignments was totally correct. The distribution of intensities of the 19 lines did not agree, and the intensity of the ^{29}Si lines was not consistent with that predicted for one ^{29}Si nucleus. That a methyl-^{13}C splitting could be

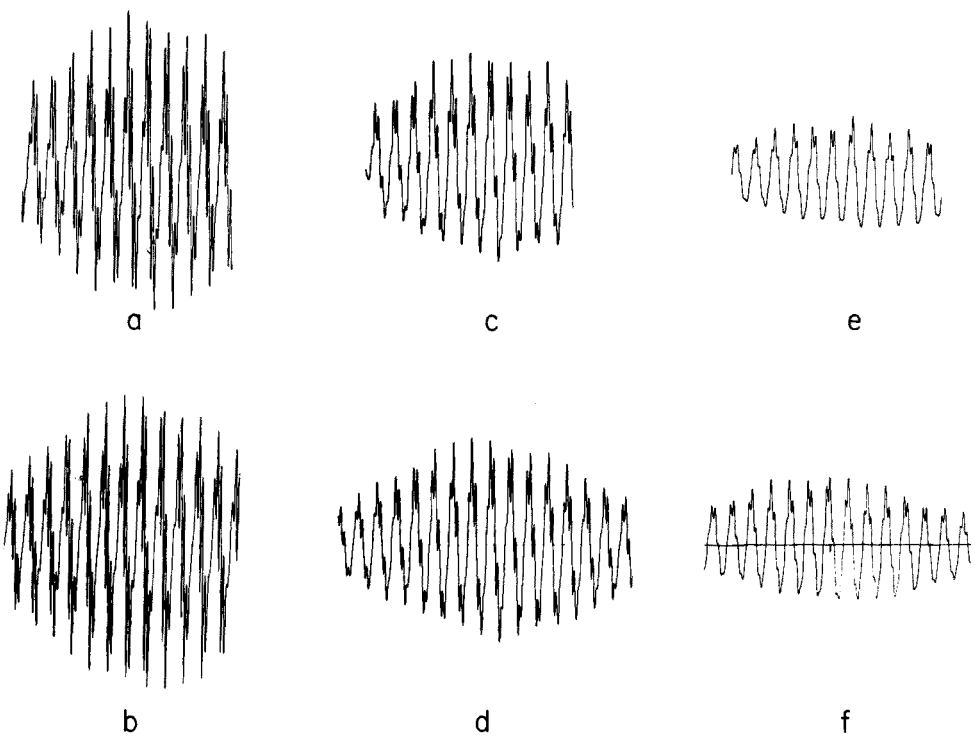


Fig. 7. Centre sections of e.s.r. spectra of the experimental (a, c, e) and computed (b, d, f) paraquat radical cation (Fig. 6) showing the effect of time constant for a time scan of 8 min. (a, b) Time constant 0.3 s; (c, d) time constant 1 s; (e, f) time constant 3 s.

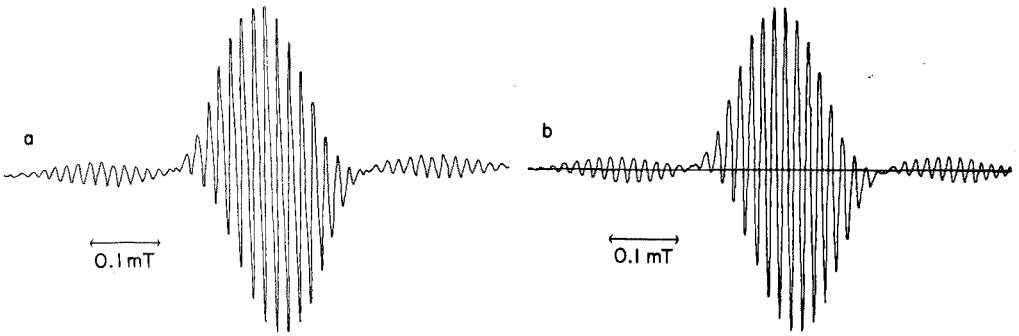


Fig. 8. (a) E.s.r. spectrum at room temperature of bis(trimethylsilyl)diacetylene radical anion in THF with K^+ as counter ion; (b) computed spectrum with the parameters found earlier [11].

involved was considered (6 methyl carbons give a "probable" abundance of ^{13}C of 6.6%); also that an alkali metal could be interacting with the radical anion. These possibilities were used in the program and the various splitting constants were varied until a good fit with the experimental spectrum obtained. Figure 8(b) shows the calculated spectrum. The method of calculation described above precludes the need to evaluate the statistical probabilities of occurrence of each of the possible combinations of the isotopic nuclei. This program calculates the spectrum as a convolution product of a series of multinomial expressions, one for each set of equivalent nuclei, without the need to expand these expressions or to neglect the smaller contributions. An alkali metal counter-ion splitting of almost double the proton splitting was also used in the reconstruction. It can be seen that a perfect fit of the experimental spectrum to the calculated spectrum was achieved (Fig. 8 a, b).

To illustrate further the power of this method of simulating the spectra arising from sets of nuclei with differing isotopes, present in natural and artificial abundances, Fig. 9 shows an absorption spectrum of a complicated

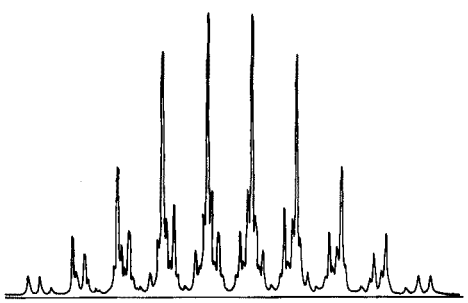


Fig. 9. Hypothetical absorption e.s.r. spectrum from three equivalent chlorine nuclei in natural isotopic abundance. (See ref. [17]).

theoretical hyperfine splitting pattern arising from three equivalent chlorine nuclei with ^{35}Cl and ^{37}Cl present in natural abundance. All that is required are the nuclear magnetogyric ratios and respective nuclear spins together with the relevant splitting constants. Reference to expression (3) shows that, in principle, the method can be extended to far more complicated cases with little loss in computation time. Figure 9 can be compared with the "stick" spectrum in the review by Hudson and Root on halogen interaction [17] which confirms the present computation.

The authors thank Professor A. G. Evans and Dr. R. P. Williams for helpful discussions. One of us (R. H. R.) thanks the S.R.C. for a C.A.S.E. studentship.

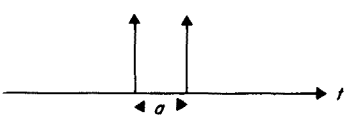
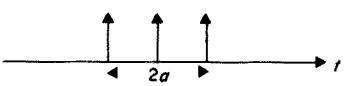
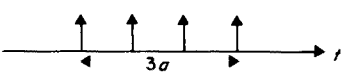
APPENDIX: FOURIER TRANSFORMS OF SPECTRAL PARAMETERS

Line shape

Time domain t	Frequency domain
$T_2/[\pi(1 + T_2^2t^2)]$ (Absorption curve; T_2 is the relaxation time)	$\exp(-2\pi f/T_2)$
n'th derivative	$(2\pi if)^n \exp(-2\pi f/T_2)$

Hyperfine coupling

Functions for a single nucleus of commonly encountered nuclear spins (where a is the coupling constant) are:

Time domain	Nuclear spin	Frequency domain
	$\frac{1}{2}$	$\cos(\pi fa)$
	1	$\frac{2}{3}(\cos(2\pi fa) + \frac{1}{2})$
	3/2	$\frac{1}{2}(\cos(\pi fa) + \cos(3\pi fa))$

and so on. It appears necessary at first to store the appropriate cosine functions for each nuclear spin case as discrete segments in the program. This inconvenience is, however, easily removed, because each set of cosine functions can be written as a polynomial in a single function $\cos(\pi fa)$. These polynomials were calculated for all known nuclear spins (see Table).

TABLE

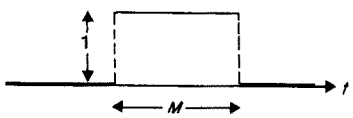
Coefficients, α_i, in the expression $\sum_i \alpha_i \cos^i(\pi f a)$ for different values of the nuclear spin. The coefficients are normalized (so that the sum of peak areas is unity) by multiplication by $2/(2I + 1)$.

I	$i = 0$	1	2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{2}$	0	1											
1	$-\frac{1}{2}$	0	2										
3/2	0	−2	0	4									
2	$\frac{1}{2}$	0	−6	0	8								
5/2	0	3	0	−16	0	16							
3	$-\frac{1}{2}$	0	12	0	−40	0	32						
7/2	0	−4	0	40	0	−96	0	64					
4	$\frac{1}{2}$	0	−20	0	120	0	−224	0	128				
9/2	0	5	0	−80	0	336	0	−512	0	256			
5	$-\frac{1}{2}$	0	30	0	−280	0	896	0	−1152	0	512		
11/2	0	−6	0	140	0	−896	0	2304	0	−2560	0	1024	
6	$\frac{1}{2}$	0	−42	0	560	0	−2688	0	5760	0	−5632	0	2048

For example, for the case $I = 1$, and for $\cos(\pi fa) = c$, then: $\cos(2\pi fa) + \frac{1}{2} = (2c^2 - 1) + \frac{1}{2} = 2c^2 - \frac{1}{2}$. These polynomials give line positions accurate to first order only. If n equivalent nuclei are present, the appropriate polynomial is raised in the nth power.

Modulation (where M *is the modulation amplitude)*

Time domain

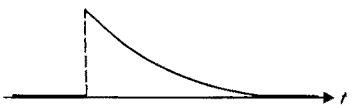


Frequency domain

$\sin(\pi fM)/\pi f$

Time constant

Time domain



$u(t)\ \exp(-t/T_c)$

Frequency domain

$T_c[1 - 2\pi ifT_c]/[(2\pi fT_c)^2 + 1]$

where $u(t)$ is the unit step function and T_c is the time constant. In the above expressions, $i = (-1)^{\frac{1}{2}}$.

In the FT multiplication of the above quantities, complex arithmetic is involved. In the program this implies the splitting of the FT into two contributions, the real and imaginary parts, which are held in separate arrays.

REFERENCES

1 D. J. E. Ingram, Free Radicals as Studied by Electron Spin Resonance, Butterworths, London, 1968, p. 121.
2 M. M. Chen, K. V. Sane, R. I. Walter and J. A. Weil, J. Phys. Chem., 65 (1961) 713.
3 E. W. Stone and A. H. Maki, J. Chem. Phys., 38 (1963) 1999.
4 R. Lefebvre and J. Maruani, J. Chem. Phys., 42 (1965) 1480, 1496.
5 D. W. Marquardt, R. G. Bennett and E. J. Burrell, J. Mol. Spectrosc., 7 (1961) 269.
6 M. Welch, Line Shape Fitting by Variable Metric Minimization, Argonne National Laboratory Program Library 1120/PhY. 220.
7 H. M. Gladney and J. D. Swalen, IBM J. Res. Develop., 8 (1964) 515.
8 See ref. 35 in J. R. Bolton and G. K. Fraenkel, J. Chem. Phys., 40 (1964) 3307.
9 A. G. Evans, J. C. Evans and E. H. Moon, J. Chem. Soc. Dalton Trans., 1974, 2390.
10 C. S. Johnson Jr. and H. S. Gutowsky, J. Chem. Phys., 39 (1963) 58.
11 A. G. Evans, J. C. Evans and C. Bevan, J. Chem. Soc. Perkin Trans. 2, (1974) 1220.
12 N. Negoita, R. Baican, F. Domsa and A. T. Balaban, Rev. Roum. Chim., 18 (1973) 6, 995.
13 R. R. Ernst, Adv. Mag. Reson., 2 (1966) 1.
14 J. W. Cooley and J. W. Tukey, Math. Computation, 19 (1965) 297.
15 E. Oran Brigham, The Fast Fourier Transform, Prentice Hall, New York, 1974.
16 A. G. Evans, J. C. Evans and M. W. Baker, J. Chem. Soc. Perkin Trans., 2 (1975) 1310.
17 A. Hudson and K. D. J. Root, Adv. Mag. Reson., 5 (1971) 1.

Analytica Chimica Acta, 103 (1978) 189—199
Computer Techniques and Optimization

OPTIMIZATION OF A WET CHEMICAL CONTINUOUS FLOW ANALYSIS METHOD EXEMPLIFIED BY THE DETERMINATIONS OF KJELDAHL NITROGEN AND TOTAL PHOSPHORUS

PETER E. ERNI* and HANS-RUDOLF MÜLLER

Swiss Federal Institute for Water Resources and Water Pollution Control (EAWAG), Dübendorf (Switzerland)

(Received 14th March 1978)

SUMMARY

The optimization of a rather complex wet chemical analysis method, such as the measurement of Kjeldahl nitrogen or total phosphorus with the Technicon AutoAnalyzer, is extremely tedious when purely empirical approaches are used. A mathematical model of the different stages of the measuring method (digestion, neutralization and color reaction) is described. The system can then be optimized for maximum measuring sensitivity. Optimization is done by solving numerically the non-linear optimization problem with constraints. The starting values for the optimization algorithm were found by varying these values systematically within the tolerated range, with checks that none of the constraints were violated. The theoretical results predict an increase in sensitivity by a factor of 15 compared to the method used previously. In practice, the sensitivity was increased by a factor of 10 for the total phosphorus method. For the simultaneous low-level determinations of Kjeldahl nitrogen and total phosphorus some problems of stability remain.

The automation of wet chemical methods of analysis includes mechanization of single basic operations as well as complex batch or continuous flow systems. Most of the wet chemical operations can be broken down to a series of dilution or enrichment steps. For complex systems, optimization with reference to maximum sensitivity, minimum sample amount, etc., can be extremely tedious if a purely empirical approach is used.

This paper describes mathematical modelling and optimization for the determination of Kjeldahl nitrogen and total phosphorus with a Technicon AutoAnalyzer, consisting of one digestion stage, one or two neutralization stages and one color reaction stage [1, 2]. The optimization is intended to maximize the concentration of nitrogen and phosphorus in the final stage (colorimetric reaction) of the process. Emphasis is placed on formulation of a mathematical model with given constraints, the optimization itself, and the application of the theoretical results to routine laboratory work.

*Present address: BBC Aktiengesellschaft Brown, Boveri und Cie, Abteilung Entsorgungstechnik XW, CH-5401 Baden/Schweiz, Werk Oerlikon.

Short description of the chemical method

The continuous stream of sample in which the nitrogen and phosphorus are to be measured, is pumped with concentrated sulfuric acid and hydrogen peroxide into the helix. The sample water is evaporated during 7 min at 320°C, then the organic compounds are decomposed and the digest is diluted. An aliquot of the diluted digest is neutralized in one or two stages, and then the reagents for the colorimetric reactions of ammonia and phosphate are added. The final measurement is made in two continuous flow-through colorimeters.

MATHEMATICAL FORMULATION OF THE MEASURING METHOD

Figure 1 shows the block diagram of the method applied for one of the measuring channels with the mathematical symbols used (see Table 1). Preliminary investigations showed that the neutralization can be done in either one or two stages (in the block diagram the second neutralization stage can be bypassed).

Mathematical model for the single-stage method

Digestion. The concentration c_1 after the digestion is given by:

$$c_1 = c_0 A / V_0' \tag{1}$$

To calculate the reduced volume after digestion (V_0')

$$V_0' = (\rho_D D + \rho_W W)/\rho(g) \tag{2}$$

it is necessary to approximate the density $\rho(g)$ and (for later use) the molarity $m_0(g)$ as a function of the $g\%$ sulfuric acid with second-order polynomials (see Tables 2 and 3). The error of the predicted $\rho(g)$ and $m_0(g)$ compared to the tabulated values is less than 1% in the applicable range of g ($27 \leqslant g \leqslant 69$, see below).

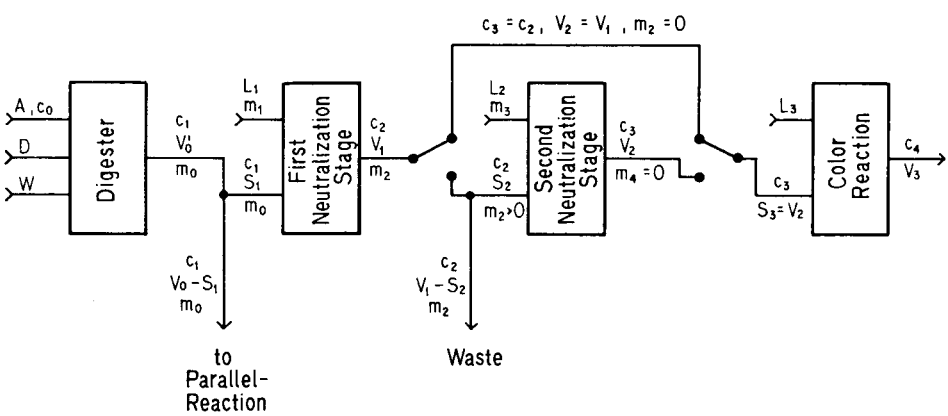


Fig. 1. Block diagram of the analytical method (see also the list of symbols).

TABLE 1

List of symbols

c_0	Normalized starting concentration of the sample (= 1 mol l^{-1})	[mol l^{-1}]
A	Volume of the sample	[ml]
D	Volume of concentrated sulfuric acid	[ml]
W	Volume of water	[ml]
V_0'	Reduced volume after digestion	[ml]
$m_0(g)$	Molarity after digestion	[mol l^{-1}]
c_1	Concentration after digestion	[mol l^{-1}]
S_1	Aliquot of V_0'	[ml]
L_1	Volume of NaOH, first neutralization stage	[ml]
m_1	Molarity of NaOH, first neutralization stage	[mol l^{-1}]
V_1	Volume after the first neutralization stage	[ml]
c_2	Concentration after the first neutralization stage	[mol l^{-1}]
m_2	Molarity of H_2SO_4 after the first neutralization stage	[mol l^{-1}]
S_2	Aliquot of V_1	[ml]
L_2	Volume of NaOH, second neutralization stage	[ml]
m_3	Molarity of NaOH, second neutralization stage	[mol l^{-1}]
V_2	Volume after the second neutralization stage	[ml]
c_3	Concentration after the second neutralization stage	[mol l^{-1}]
L_3	Sum of reagents for the color reaction	[ml]
S_3	Aliquot of V_2 (= V_2)	[ml]
V_3	Volume after the color reaction	[ml]
c_4	End concentration	[mol l^{-1}]
$\rho(g)$	Density of g% sulfuric acid	[g ml^{-1}]
ρ_D	Density of concentrated sulfuric acid	[g ml^{-1}]
ρ_W	Density of water	[g ml^{-1}]
g	Percentage (by weight) of sulfuric acid after digestion	[—]
V_F	Carrier volume of the digestion multiplied by time	[ml min]
r	Revolutions per min of the helix	[l min^{-1}]
s	Experimentally obtained solubility for Na_2SO_4 in H_2SO_4 (= 2.9 mol l^{-1})	[mol l^{-1}]
m_m	Maximum molarity of commercially available NaOH (= 10.8 mol l^{-1})	[mol l^{-1}]

The ranges for the volumes of sample, sulfuric acid and water, which are limited by the characteristics of the apparatus, give the constraints N6—N8 (Tables 2 and 3). In addition, the sums of the volumes of A and D, D and W, respectively, are restricted by the capacity of the helix, giving the constraints N1 and N2 (Tables 2 and 3).

Neutralization. If the reduction in volume is neglected, V_1 is given by

$$V_1 = S_1 + L_1 \tag{3}$$

The concentration c_2 after the neutralization is given by

$$c_2 = c_1 S_1 / V_1 \tag{4}$$

The ranges for the volumes for S_1 and L_1 give the constraints N9 and N10 (Tables 2 and 3).

TABLE 2

Non-linear optimization problem for the single-stage method

Performance index

$$c_4 = \frac{c_0 A S_1 \rho(g)}{(\rho_D D + \rho_W W)(S_1 + L_1 + L_3)} = \text{max.} \qquad \text{Z1}$$

$$g = [\rho_D D/(\rho_D D + \rho_W W)]100$$

Constraints

$D + A - V_F r \leqslant 0$	N1
$D + W - V_F r \leqslant 0$	N2
$2S_1\rho(g) - \rho_D D - \rho_W W \leqslant 0$	N3
$S_1 m_0(g) - sL_1 - sS_1 \leqslant 0$	N4
$2S_1 m_0(g) - m_m L_1 \leqslant 0$	N5
$2.0 \leqslant A \leqslant 5.0$	N6
$1.0 \leqslant D \leqslant 2.5$	N7
$2.0 \leqslant W \leqslant 5.0$	N8
$0.4 \leqslant S_1 \leqslant 2.0$	N9
$0.1 \leqslant L_1 \leqslant 1.0$	N10
$0.1 \leqslant L_3 \leqslant 2.0$	N11

Values to be determined

A, D, W, S_1, L_1, L_3

Constants

$r = 4\ \text{l min}^{-1}$	$V_F = 2\ \text{ml min}$
$\rho_D = 1.84\ \text{g ml}^{-1}$	$s = 2.9\ \text{mol l}^{-1}$
$c_0 = 1\ \text{mol l}^{-1}$	$m_m = 10.8\ \text{mol l}^{-1}$
$\rho_W = 1.0\ \text{g ml}^{-1}$	

Polynomials for $\rho(g)$ and $m_0(g)$

$$\rho(g) = 1.02351 + 5.0851 \times 10^{-3} g + 4.71179 \times 10^{-5} g^2$$
$$m_0(g) = 4.64268 \times 10^{-1} + 7.24837 \times 10^{-2} g + 1.21151 \times 10^{-3} g^2$$

Another restriction for S_1 is $S_1 \leqslant 0.5 V_0'$, because V_0' must be split between two measuring channels. The condition for neutrality is given by $m_1 L_1 = 2m_0(g)S_1$; the limit of solubility for sodium sulfate is given by $m_1 L_1 \leqslant 2sV_1$; and the molarity of sodium hydroxide is given by $m_1 \leqslant m_m$.

Color reaction. After the color reaction, the volume V_3 is given by

$$V_3 = S_3 + L_3 \tag{5}$$

For the single-stage method $S_3 = V_1$ and $c_3 = c_2$. The end concentration c_4 after the color reaction is given by

$$c_4 = c_3 S_3 / V_3 \tag{6}$$

The range for the volume L_3 is given by constraint N11 in Tables 2 and 3.

TABLE 3

Non-linear optimization problem for the two-stage method. The constants and the polynomials for $\rho(g)$ and $m_0(g)$ are the same as in Table 2

Performance index

$$c_4 = \frac{c_0 A S_1 S_2 \rho(g)}{(\rho_D D + \rho_W W)(S_1 + L_1)(S_2 + L_2 + L_3)} = \text{max.} \qquad \text{Z2}$$

$$g = [\rho_D D/(\rho_D D + \rho_W W)]100$$

Constraints

$D + A - V_F r \leqslant 0$	N1
$D + W - V_F r \leqslant 0$	N2
$2S_1\rho(g) - \rho_D D - \rho_W W \leqslant 0$	N3
$S_2 - 0.7\,S_1 - 0.7\,L_1 \leqslant 0$	N12
$2S_1S_2m_0(g) - m_m S_1 L_2 - m_m L_1 L_2 - m_m L_1 S_2 \leqslant 0$	N13
$S_1L_1S_2m_0(g) - sS_1^2L_2 - 2sS_1L_1L_2 - sL_1^2L_2 - sL_1^2S_2 - sS_1L_1S_2 \leqslant 0$	N14
$S_1S_2m_0(g) - sL_1S_2 - sS_1L_2 - sL_1L_2 - sS_1S_2 \leqslant 0$	N15
$2.0 \leqslant A \leqslant 5.0$	N6
$1.0 \leqslant D \leqslant 2.5$	N7
$2.0 \leqslant W \leqslant 5.0$	N8
$0.4 \leqslant S_1 \leqslant 2.0$	N9
$0.1 \leqslant L_1 \leqslant 1.0$	N10
$0.4 \leqslant S_2 \leqslant 2.0$	N16
$0.4 \leqslant L_2 \leqslant 2.0$	N17
$0.1 \leqslant L_3 \leqslant 2.0$	N11

Values to be determined

$A, D, W, S_1, L_1, S_2, L_2, L_3$

By insertion and transformation of the above equations and inequalities, one obtains the performance index Z1 and 11 constraints (N1—N11) for the optimization problem (Table 2).

Mathematical model for the two-stage method

Digestion. Changes from the single-stage method are not required. Equations (1) and (2) and constraints N1, N2, N6, N7 and N8 are also valid for the two-stage method.

First neutralization stage. If the neutralization is done in two stages, with the first stage giving only partial neutralization, the equation $m_1L_1 = 2m_0(g)S_1$ is replaced by $m_2 > 0$, and

$$m_2 = (2S_1m_0(g) - L_1m_1)/V_1 \qquad (7)$$

The other equations and inequalities still apply.

Second neutralization stage. In this stage the residual sulfuric acid becomes completely neutralized; it is assumed that the sodium hydroxide has the same molarity as in the first stage ($m_3 = m_1$).

The volume V_2 after the second neutralization stage is given by:

$$V_2 = S_2 + L_2 \tag{8}$$

The concentration c_3 after the second neutralization stage is given by:

$$c_3 = c_2 S_2 / V_2 \tag{9}$$

The ranges for the volumes S_2 and L_2 are given by the constraints N16 and N17 in Table 3. Another restriction for S_2 is $S_2 \leqslant 0.7 V_1$, which is needed because the air which delimits the liquid segments must be removed by a partial stream before passing the pump a second time. The condition for neutrality is given by $m_1 L_2 = m_2 S_2$, and the limit of solubility for sodium sulfate is given by $L_1 m_1 S_2 + L_2 m_3 V_1 \leqslant 2 s V_1 V_2$.

Color reaction. The equations and inequalities are the same as for the single-stage method, but $S_3 = V_2$.

By insertion and transformation of the appropriate equations and inequalities, one obtains the performance index Z2 and 15 constraints (N1—N3, N6—N17) of the optimization problem; these are tabulated in Table 3.

THE OPTIMIZATION METHOD

As can be seen from Tables 2 and 3, 6 variables must be determined for the single-stage method and 8 variables for the two-stage method. The performance indices Z1 and Z2 indicate that A should be made as high as possible, because this variable occurs only in the nominator of Z1 and Z2, and the inequalities N1 and N2 are satisfied for all possible values of A, D and W (given by N6—N8). From Z1 and Z2 it can also be deduced that L_3 should be as small as possible, because this value occurs only in the denominator of Z1 and Z2 and in no inequality. With these considerations, the optimization problem is reduced to the optimization of 4 or 6 variables.

In general, the result of an optimization problem is one set of values calculated to be optimal. However, since in practice, one certainly may use other than the optimal values, it is more convenient to calculate the optimal end concentration c_4 as a function of g, the percentage concentration of sulfuric acid. Preliminary investigations showed that g may vary between 27 (for $D = 1$ and $W = 4.97$) and 64 (for $D = 1.93$ and $W = 2$) for the single-stage method, and between 27 (for $D = 1$ and $W = 4.97$) and 69 (for $D = 2.42$ and $W = 2$) for the two-stage method.

The non-linear optimization problem with constraints was solved with the aid of a computer program package described by Rufer [3]. The transformation of the non-linear optimization problem with constraints into one without constraints was done by the method of Kelly et al. (accelerated exterior point method) [4]. The unconstrained optimization problem was solved by the method of Fletcher [5], and the search for the one-dimensional optimum was done by golden section search with an algorithm to bracket a starting interval [6]. The starting values for the optimization method were

found by varying these values systematically in the tolerated range in such a way that none of the inequalities was violated.

When the program package of Rufer was used, the optimum could be found only when the initial values (especially D, W and L_1) were rather close to the values found by varying systematically these values in the tolerated range. Furthermore, the optima found by the optimization program package for the two-stage method were only slightly better than those obtained by varying systematically the values to be determined within their tolerated range. For the two-stage method the optima were not very pronounced, so that the optimal values L_1, S_2 and L_2 could vary considerably without significantly affecting the final concentration c_4.

THEORETICAL RESULTS OF THE OPTIMIZATION

The numerical results for the single-stage method are tabulated in Table 4. The highest final concentration c_4 occurs with a 45—47% concentration of sulfuric acid. The optimal values are (for $g = 46$):

$A = 5.0$ ml; $D = 1.0$ ml; $W = 2.34$ ml; $S_1 = 0.9$ ml; $L_1 = 1.0$ ml; $L_3 = 0.1$ ml.

With these values the final concentration is $c_4 = 0.734$ mol l^{-1} if the starting concentration is $c_0 = 1$ mol l^{-1}.

For the two-stage method, the results are tabulated in Table 5. The highest final concentration c_4 again occurs at a 45—47% concentration of sulfuric acid. The optimal values are (for $g = 46$):

$A = 5.0$ ml; $D = 1.0$ ml; $W = 2.16$ ml; $S_1 = 1.47$ ml; $L_1 = 0.4$ ml; $S_2 = 1.31$ ml; $L_2 = 0.95$ ml; $L_3 = 0.1$ ml.

With these values the final concentration is $c_4 = 0.741$ mol l^{-1} if the starting concentration is $c_0 = 1$ mol l^{-1}.

The final concentrations c_4 for the single-stage and two-stage methods are plotted as a function of g in Fig. 2. It can be seen that the two-stage method gives slightly better results for $g > 31$ than the single-stage method. The superiority of the two-stage method increases with increasing concentration of sulfuric acid. The curves of the two methods are not symmetrical. This asymmetry may be useful, if sub-optimal values are required for practical reasons.

The values used initially here [1] to measure nitrogen ($A = 2$ ml, $D = 2.2$ ml, $W = 2.0$ ml, $S_1 = 0.42$ ml, $L_1 = 0.8$ ml, $S_2 = 0.42$ ml, $L_2 = 0.8$ ml) are extremely bad. With these values the range is such that only the two-stage method can be used and, even with optimal values of S_1, L_1, S_2 and L_2, only a low end concentration c_4 is obtained. The values chosen by El Kei [2], to measure nitrogen ($A = 4$ ml, $D = 2.2$ ml, $W = 5.0$ ml, $S_1 = 1.0$ ml, $L_1 = 0.6$ ml, $S_2 = 1.2$ ml, $L_2 = 0.42$ ml) are much better. With these values a value for g is obtained where the curve has its maximum, although the values are not optimal.

These theoretical investigations allow the following conclusions to be

TABLE 4

Numerical results for the single-stage method

g (—)	$\rho(g)$ (g ml^{-1})	$m_0(g)$ (mol l^{-1})	D (ml)	W (ml)	S_1 (ml)	L_1 (ml)	c_4 (mol l^{-1})	m_1 (mol l^{-1})
27	1.20	3.30	1.00	4.97	1.64	1.00	0.525	10.8
31	1.23	3.88	1.00	4.09	1.39	1.00	0.578	10.8
35	1.26	4.49	1.00	3.41	1.20	1.00	0.626	10.8
39	1.29	5.13	1.00	2.88	1.05	1.00	0.671	10.8
44	1.34	6.00	1.00	2.34	0.90	1.00	0.721	10.8
45	1.35	6.18	1.00	2.25	0.87	1.00	0.730	10.8
46	1.36	6.36	1.00	2.16	0.84	1.00	0.734	10.7
47	1.37	6.55	1.00	2.07	0.80	1.00	0.733	10.4
48	1.38	6.73	1.00	2.00	0.76	1.00	0.729	10.2
52	1.42	7.51	1.18	2.00	0.63	1.00	0.618	9.5
56	1.46	8.32	1.38	2.00	0.54	1.00	0.524	8.9
60	1.50	9.17	1.63	2.00	0.46	1.00	0.444	8.5
64	1.54	10.07	1.93	2.00	0.41	1.00	0.374	8.2

TABLE 5

Numerical results for the two-stage method

g (—)	$\rho(g)$ (g ml^{-1})	$m_0(g)$ (mol l^{-1})	D (ml)	W (ml)	S_1 (ml)	L_1 (ml)	S_2 (ml)	L_2 (ml)	c_4 (mol l^{-1})	m_1 (mol l^{-1})
27	1.20	3.30	1.00	4.97	2.00	0.40	1.68	0.58	0.521	10.8
31	1.23	3.88	1.00	4.10	2.00	0.40	1.68	0.72	0.578	10.8
35	1.26	4.49	1.00	3.42	2.00	0.40	1.68	0.88	0.630	10.8
39	1.29	5.13	1.00	2.88	1.82	0.40	1.56	0.93	0.676	10.8
44	1.34	6.00	1.00	2.34	1.56	0.40	1.37	0.93	0.727	10.8
45	1.35	6.18	1.00	2.25	1.52	0.40	1.34	0.93	0.737	10.8
46	1.36	6.36	1.00	2.16	1.47	0.40	1.31	0.95	0.741	10.7
47	1.37	6.55	1.00	2.07	1.43	0.40	1.28	0.98	0.740	10.4
48	1.38	6.73	1.00	2.00	1.40	0.40	1.26	1.01	0.738	10.2
53	1.43	7.71	1.23	2.00	1.49	0.40	1.32	1.45	0.608	9.3
57	1.47	8.53	1.44	2.00	1.59	0.72	1.61	1.65	0.520	8.8
61	1.51	9.39	1.70	2.00	1.70	1.00	1.89	1.96	0.443	8.4
65	1.55	10.29	2.02	2.00	1.84	1.00	1.54	2.00	0.372	8.1
69	1.60	11.23	2.42	2.00	2.00	1.00	1.26	2.00	0.310	7.8

reached. It is possible to determine Kjeldahl nitrogen and total phosphorus with either the single-stage method or the two-stage method. The final concentration c_4 is only slightly smaller in the former method. Secondly, if for practical reasons, the optimal values cannot be used, it is advisable to go in the direction of lower concentration of sulfuric acid, because then the final concentration will decrease more gradually. Thirdly, in the two-stage method, L_1 influences the final concentration c_4 only if $g \geqslant 57$. For $g < 57$, L_1 can be varied extensively, without change of c_4. However, the values of S_1, L_2 and S_2 will change if L_1 is varied. This advantage of the two-stage method can be

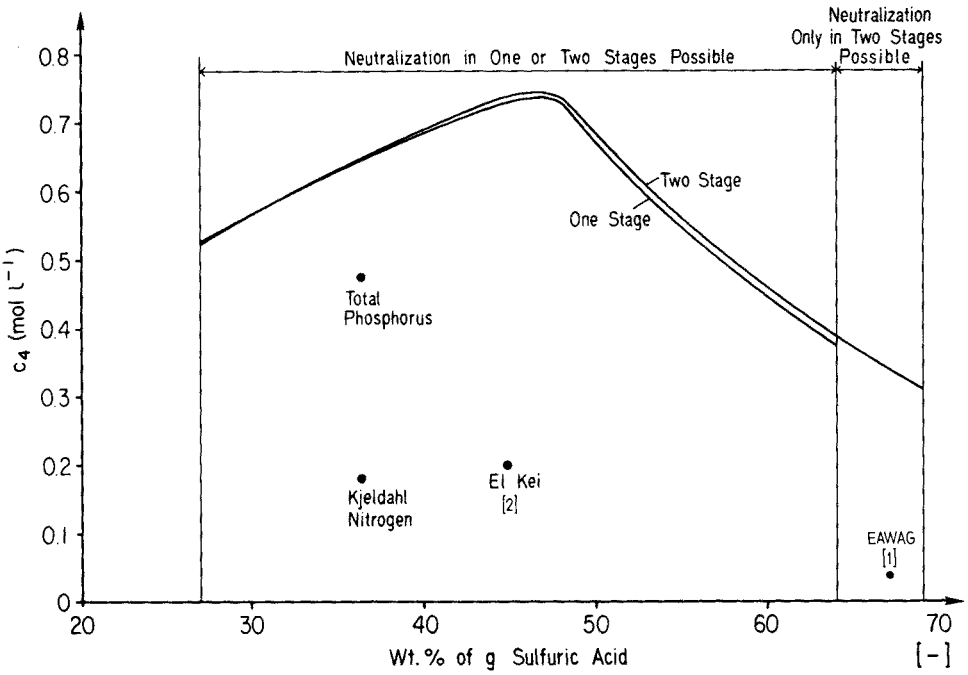


Fig. 2. Optimal final concentration c_4 as a function of g for the single- and two-stage methods.

of practical value, as with commercially available peristaltic tubes, it is not possible to pump an arbitrary volume, and so some discrete L_1 value must be used.

APPLICATION OF THE THEORETICAL RESULTS

With a single-channel AutoAnalyzer, the sensitivity and stability for both the one-stage and the two-stage methods were tested separately for nitrogen and phosphorus on the basis of the calculated results. Starting values for the volumes of reagents were the calculated values which lie nearest to the given pumping capacity of the peristaltic tube. The sensitivity was checked by comparing the difference of absorbance between twice-distilled water and standard solutions containing either 500 μg N l^{-1} or 500 μg P l^{-1} with normalized instrument adjustment (standard calibration 1). The critical basicity (nitrogen) or acidity (phosphorus) which in effect determines the respective color reaction, was checked by titration of the overflow of the colorimeter with 0.1 M acid or 0.1 M base [1]. If necessary, the sodium hydroxide concentration of the reagent volume L_1 (one-stage method) or L_2 (two-stage method) was adjusted.

The increase in sensitivity obtained for all four experimental cases is in

the range predicted by theory. Because of hydraulic and chemical stability, the two-stage method is to be preferred for either measurement.

The following facts caused some problems and necessitated changes compared to the calculated values. First, the transport of the digesting acid in the helix is not constant when D = 1.0 ml, and the helix is not completely moistened. Thus D had to be increased to D = 1.3 ml. Secondly, the temperature program 320—300—200°C gives incomplete evaporation of the sample water. This results in inhomogeneous concentration of the digest, residual hydrogen peroxide and reduced efficiency of digestion for organic nitrogen. A new temperature program 360—340—300°C and a flow of only 0.05 ml min^{-1} for the 32% hydrogen peroxide were chosen. The (possibly) remaining hydrogen peroxide could be reduced by adding sodium hydrogensulfite to the reagent volume L_2.

Finally, even with the increased temperature program, the digestion of organic nitrogen is incomplete for those compounds having high bond energy. Modifications to the digester, e.g. reduction of the sample volume A, have a direct effect on the sensitivity of both measurement channels.

For the simultaneous low-level determinations of Kjeldahl nitrogen and total phosphorus, additional investigations are necessary.

Operating instructions

Setting up the system should be done as described by the manufacturer.

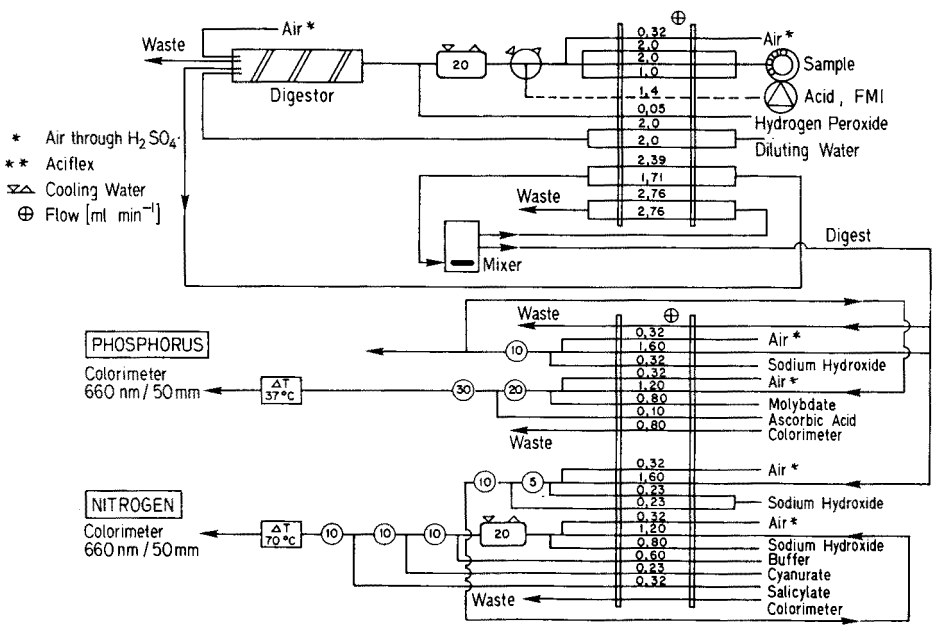


Fig. 3. Flow diagram for the determination of Kjeldahl nitrogen and total phosphorus.

The following points in the flow diagram (Fig. 3) need special attention:

(a) for measuring unfiltered river-water samples, stirring during sampling is necessary;
(b) the sulfuric acid is measured and added by an FMI—piston pump;
(c) the whole digestion unit should have an inclination of 1% towards the outflow;
(d) an air bubbler should be installed to provide better mixing of the digested sample with the diluting water within the turns of the helix.

Finally, the mixing chamber recommended by El Kei [2] was inserted between the digestion unit and the neutralization stage. The disadvantage of the mixer is a slightly higher memory effect. The reagents used are listed in Table 6.

TABLE 6

Reagents

Kjeldahl nitrogen			*Total phosphorus*		
L_1	Sodium hydroxide	32%	L_1	Sodium hydroxide	32%
L_2	Sodium hydroxide	24%	L_2	Sodium hydroxide	18%
	Aerosol 22	2 drops		Molybdate	1%
L_3'	Trisodium phosphate	5%		KSb tartrate	0.1%
	NaK tartrate	5%		Aerosol 22	2 drops
	Aerosol 22	2 drops	L_3	Ascorbic acid	1.5%
L_3''	Cyanurate	0.4%		Acetone	5%
	Sodium hydroxide	7.5%			
L_3'''	Salicylate	20%			
	Sodium nitroprusside	0.12%			

(L_3', L_3'' and L_3''' form the L_3 used in the mathematical model)

REFERENCES

1 H. R. Müller and J. Zobrist, Die automatische und simultane Bestimmung von Kjeldahl-Stickstoff und Gesamt-Phosphor in Oberflächengewässern; unpublished EAWAG report.
2 O. El Kei, Anal. Chim. Acta, 86 (1976) 63.
3 D. F. Rufer, General Purpose Nonlinear Programming Package, Fachbericht Nr. 77-02, Fachgruppe für Automatik, Eidgenössische Technische Hochschule Zürich, 1977.
4 H. J. Kelley, W. F. Denham, I. L. Johnson and P. O. Wheatley, J. Astronaut. Sci., 13 (1966) 166.
5 R. Fletcher, Comput. J., 13 (1970) 317.
6 D. F. Rufer, Dissertation No. 5519, Eidgenössische Technische Hochschule Zürich, 1975.

VOL. 103, No. 2 ANALYTICA CHIMICA ACTA June 15, 1978

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

CONTENTS

Application of a text search system based on Boolean strategy to mass spectral data identification
J. A. de Haseth, H. B. Woodruff, S. L. Lowry and T. L. Isenhour (Chapel Hill, NC, U.S.A.) . . 109
An approach to automated partial structure expansion
C. A. Shelley, T. R. Hays, M. E. Munk and R. V. Roman (Tempe, AZ, U.S.A.) 121
Missing values in time series and the implications on autocorrelation analysis
C. B. G. Limonard (Nijmegen, The Netherlands) 133
A substructure-oriented ^{13}C-n.m.r. chemical shift retrieval system
J. Zupan, S. R. Heller (Washington, D.C., U.S.A.), G. W. A. Milne (Bethseda, MA, U.S.A.) and
J. A. Miller (Towson, MD, U.S.A.) 141
The learning machine in quantitative chemical analysis. Part 1. Anodic stripping voltammetry of cadmium, lead and thallium
M. Bos and G. Jasink (Enschede, The Netherlands) 151
Quantitative analysis for polycyclic aromatic hydrocarbons by spectral decomposition of molecular fluorescence
H. S. Gold, C. E. Rechsteiner, Jr. and R. S. Buck (Chapel Hill, NC, U.S.A.) 167
Simulation of electron spin resonance spectra by fast Fourier transform. A novel method of calculating spectra to include isotopic substitution, superhyperfine coupling, instrument time constant and modulation broadening in fluid and polycrystalline media
J. C. Evans, P. H. Morgan and R. H. Renaud (Cardiff, Gt. Britain) 175
Optimization of a wet chemical continuous flow analysis method exemplified by the determinations of Kjeldahl nitrogen and total phosphorus
P. E. Erni and H.-R. Müller (Dübendorf, Switzerland) 189

Printed in The Netherlands