

# **ANALYTICA CHIMICA ACTA**

International journal devoted to all branches of analytical chemistry

## **COMPUTER TECHNIQUES AND OPTIMIZATION**

EDITOR

J. T. CLERC (Bern, Switzerland)

Associate Editor

E. ZIEGLER (Mülheim, Germany)

Editorial Advisers

R. E. Dessy, Blacksburg, VA

J. W. Frazer, Livermore, CA

H. Günzler, Ludwigshafen

S. R. Heller, Washington, DC

Z. Hippe, Rzeszów

J. F. K. Huber, Vienna

T. L. Isenhour, Chapel Hill, NC

P. C. Jurs, University Park, PA

D. L. Massart, Sint Genesius-Rhode

S. Sasaki, Toyohashi

H. C. Smit, Amsterdam

# 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 1981 (incorporating the section on Computer Techniques and Optimization)

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	123	124/1	124/2	125	126	127	128	129	130/1	130/2	131	132
Section on Computer Techniques and Optimization		133/1			133/2			133/3			133/4	

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

**Submission of Papers.** Manuscripts (three copies) should be submitted as designated below for rapid and efficient handling:

*Papers from the Americas to:* Professor Harry L. Pardue, Department of Chemistry, Purdue University, West Lafayette, IN 47907, U.S.A.

*Papers from all other countries to:* Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box Birmingham B15 2TT, England.

For the section on *Computer Techniques and Optimization:* Dr. J. T. Clerc, Universität Bern, Pharmazeutisches Institut, Sahlstrasse 10, CH-3012 Bern, Switzerland.

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

Submission of an article is understood to imply that the article is original and unpublished and is not being considered for publication elsewhere. Upon acceptance of an article by the journal, the author(s) resident in the U.S.A. are asked to transfer the copyright of the article to the publisher. This transfer will ensure the widest dissemination of information under the U.S. Copyright Law.

**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. 121, p. 353 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 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 11 volumes in 1981. The subscription for 1981 (Vols. 123-133) is Dfl. 1639.00 plus Dfl. 198.00 (postage) approx. U.S. \$942.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 133) 149.00 plus Dfl. 18.00 (postage) (total approx. U.S. \$86.00). Journals are sent automatically by airmail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All 1981 volumes (Vols. 1-121) except Vols. 23 and 28 are available at Dfl. 164.00 (U.S. \$84.00), plus Dfl. 13.00 (U.S. \$) postage and handling, per volume.

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

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

Proceedings of the  
Fifth International Conference on Computers in  
Chemical Research and Education

*Toyohashi, Japan, October 15–17, 1980*

#### PUBLISHER'S NOTE

This issue of *Computer Techniques and Optimization* is the last one to appear. Separate publication will no longer take place because papers on these topics are now an integral part of analytical chemistry and will henceforth be published in the regular issues of *Analytica Chimica Acta*.

A detailed explanation of this development appears as an Editorial in Volume 134, No. 1, the January 1982 issue of the journal.



## THE DENDRAL PROJECT: RECENT ADVANCES IN COMPUTER-ASSISTED STRUCTURE ELUCIDATION<sup>†</sup>

DENNIS H. SMITH\*, NEIL A. B. GRAY, JAMES G. NOURSE and CHRISTOPHER W. CRANDELL

*Department of Chemistry, Stanford University, Stanford, California 94305 (U.S.A.)*

(Received 23rd January 1981)

### SUMMARY

Recent advances in the DENDRAL Project on computer-assisted approaches to structure elucidation are reviewed. Important developments include: (1) novel approaches to constrained generation of constitutional isomers to obtain structural candidates for an unknown; (2) the first solution to the problem of exhaustive generation of configurational stereoisomers and the implementation of a program for constrained generation of such stereoisomers; (3) the refinement of methods for predicting mass spectra of molecular structures and for using such predicted spectra to determine the relative plausibilities of different hypothesized structures for an unknown; (4) the development of experimental approaches for the interpretation and prediction of <sup>13</sup>C-n.m.r. spectra. Future extensions allowing the representation and use of conformational stereochemical information, and potential applications of such extensions are discussed.

Research within the DENDRAL Project into computer-assisted methods for chemical structure elucidation is now well into its second decade. This research originated in studies on methods for generating acyclic structures, and for characterizing cyclic structures [1]. The name DENDRAL for this project is an acronym derived from the name of the first algorithm developed — the DENDRitic ALgorithm for acyclic isomer generation. A significant aspect of this early analysis of structural forms was that it sought to guarantee completeness in the sense that structure generation procedures could be made exhaustive and irredundant. The exhaustiveness of a structure generation algorithm is an important aspect of computer-aided structural analysis, for it alone can guarantee that all possible structures are considered for an unknown.

Even for simple acyclic structures, the number of isomers for a given molecular composition can be extremely large [2]. It was evident that it would be impractical to attempt to develop an approach to structure elucidation that relied solely on the generation of possible isomeric structures followed by testing for compatibility with available chemical and spectral

<sup>†</sup>Part XXXVIII of the series Applications of Artificial Intelligence for Chemical Inference. For Part XXXVII see R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse and C. Djerassi, *J. Org. Chem.*, 46 (1981) 1708.

data. The idea of using inferred substructural constraints became an intrinsic part of the DENDRAL approach [3]. These inferred substructural constraints would guide an exhaustive structure generator so that only a small subset of plausible structures would be created and tested against the complete spectral data. Thus, the approach to structure elucidation illustrated in Fig. 1 was derived. This approach involves three distinct phases; (1) a PLANNING phase in which substructural constraints are inferred from available spectral and chemical data; (2) a GENERATION phase in which all structures compatible with these substructural constraints are created; (3) a TEST phase in which the generated candidate structures are re-evaluated in the light of all available data with some candidates being eliminated and the remainder being rank-ordered according to some measure of their consistency with the observed data.

Initial work on the Heuristic DENDRAL program led to a PRELIMINARY INFERENCE MAKER that could derive substructural constraints through a simple analysis of mass spectra and could utilize certain  $^1\text{H-n.m.r.}$  data [4, 5]. The substructural constraints derived by the PRELIMINARY INFERENCE MAKER controlled the DENDRAL acyclic isomer generator. Functions were also developed to predict the mass spectral fragmentations of simple acyclic monofunctional molecules; comparison of predicted and observed spectral data permitted the ranking of generated candidates [6]. Although its scope was limited by the restrictions of the structure generator and of the approaches to interpreting and predicting spectra, this Heuristic DENDRAL system constituted the first fully automated approach to structure analysis.

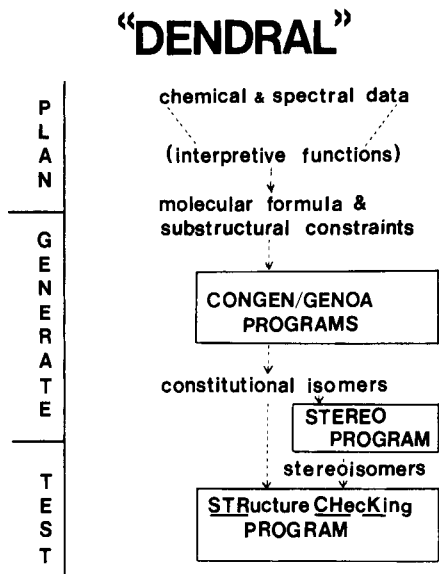


Fig. 1. The DENDRAL approach to computer-assisted structure elucidation.

Structures of chemical interest are for the most part fairly elaborate polyfunctional, polycyclic systems. Most of the work conducted on the DENDRAL project throughout the 1970's has been concerned with the development of methods for generating and manipulating representations of polycyclic systems. Based on mathematically proven algorithms, a systematic method of identification of all possible structural isomers consistent with a given molecular formula was devised [7]. These isomer generating functions were subsequently incorporated into a complete system, CONGEN (for CONstrained GENERator), that permitted the investigator to define substructural constraints that could limit the generation process [8]. CONGEN allows for the use of arbitrarily elaborate "superatom" constituents in a structure; these superatoms represented distinct subparts of the unknown molecule whose presence could be unambiguously established from available spectral or chemical data. In addition to the superatom building blocks, CONGEN allowed the investigator to define additional constraints on the bonding between different atoms and/or superatoms; these additional constraints could specify desired features (GOODLIST) or undesired combinations (BADLIST). CONGEN has been applied to a variety of problems of structure elucidation in various laboratories [9]. A second version of the CONGEN program, employing a rather different approach to structure generation, was subsequently developed. This version is in use in about twenty industrial and academic research laboratories.

CONGEN provides a powerful aid to determining the number and nature of possible structures consistent with the available chemical and spectral data characterizing an unknown. However, the CONGEN approach has some limitations. The distinction between superatoms and GOODLIST constraints requires that an investigator consider in detail the possible implications of each separate piece of spectral evidence. The fact that the program works only in terms of molecular constitution, ignoring stereoisomerism, is a further limitation meaning that "structures" derived by the program are intrinsically ambiguous. Further, CONGEN corresponds only to the GENERATE phase of the DENDRAL model of structure elucidation. CONGEN does include experimental modules that enable an investigator to survey a set of generated candidates to determine if any of the structures incorporate previously identified standard skeletons [10], or to estimate the compatibility of a candidate structure with observed mass spectral data [11, 12]. However, the testing and evaluation of generated candidates is largely left to the investigator. The CONGEN program itself provides no aid for the initial processes of inferring substructural constraints from spectral data.

Many of these limitations of the CONGEN system have been addressed in recent work. A new structure generator, GENOA, provides much greater flexibility in the forms allowed for substructural constraints. Constitutional isomers resulting from CONGEN and GENOA can now be processed further to derive all possible configurational stereoisomers compatible with known stereochemical constraints. The functions for testing the compatibility of

candidate structures with mass spectral data have been further refined and similar functions for analysis of  $^{13}\text{C}$ -n.m.r. spectral data have also been developed. Of a more experimental nature are functions that attempt to assist the investigator in inferring complex substructural constraints from  $^{13}\text{C}$ -n.m.r. data. These constraints are then employed to guide the structure generation processes.

In subsequent sections these developments are described. Much of the detailed description of algorithms and computer programs is to be found in the references to recent publications. In this paper, the methods are introduced and a few examples of applications to current structural problems are presented.

#### GENERATION OF CONSTITUTIONAL ISOMERS: THE GENOA PROGRAM

In the last decade, many programs for structure/isomer generation have been presented. In addition to the CONGEN program [8], there are the CASE system of Munk and co-workers [13], the CHEMICS system of Sasaki and co-workers [14] and the STREC system of Gribov et al. [15]. These methods all have important limitations in their ability to handle structure generation problems involving complex structures [16]. This is due in part to the requirement for obtaining (manually or automatically) substructural inferences from "real-world" spectral and chemical data. These data are unfortunately characterized by being both ambiguous (one or more alternative substructures for a given spectral or chemical signature) and redundant (highly overlapping substructures).

Manual approaches to structure elucidation must keep substructural alternatives in mind in order to cope with ambiguity. However, the standard programs for structure generation have no direct mechanisms for dealing with substructural alternatives. Each alternative interpretation must be treated as a separate structure elucidation problem. The redundancy of spectral data also must be kept in mind in sorting out substructural inferences obtained from a variety of data because of the potential for overlapping partial structures (substructures or molecular fragments). Current computer programs have either no or only partial [13] mechanisms for treating such overlaps; in the terminology of CONGEN, all superatoms must be completely disjoint.

A new structure generator, GENOA, has recently been completed [16] to cope with such limitations and to emulate more closely manual approaches to structure elucidation. GENOA is capable of taking into account all possible overlaps of substructures and all alternative substructures at the very beginning of the computational procedure. GENOA solves a structural problem using an algorithm called "constructive substructure search" in which different substructures are pieced together with automatic consideration of potential overlapping and specified alternative substructures.

The use of GENOA is illustrated here by choosing from the literature a

recent problem described by the authors as involving several overlapping substructures. As before [16] the structural information is given to GENOA in the approximate order in which it is presented in the paper. This illustrates the utility of GENOA as used during the course of elucidation of the structure of a new compound.

Before the example is discussed, several features of the GENOA program [16] must be noted, in order that the presentation which follows be more clear. Substructural constraints are supplied to the program one at a time, together with a specified range of occurrence, e.g., at least two, exactly 4, none, ... of a specific substructure. GENOA incorporates the required number using a constructive substructure search algorithm [16] which explicitly considers possible overlaps of the newly specified substructure with structural components already built into the problem from previous substructural constraints. The results of incorporation of new constraints are called "cases" [16]. A case consists usually of a set of disconnected substructures resulting from incorporation of one or more previous constraints together with several atoms as yet unassigned to specific parts of the structure. As more and more constraints are specified, the cases are elaborated further into more complete structures as additional atom interconnections are formed. GENOA never constructs the complete set of structural isomers which may be obtained from a case or cases until requested to do so using the GENERATE command [16]. This serves to keep the size of the problem in terms of numbers of cases manageable throughout a problem. Although early in a problem (when few constraints have been specified) there may be millions of possible structures, an investigator deals with the relatively few cases which specify only what is known about the structure at that point.

The degree of overlap of substructures can be controlled by the investigator in situations where information on the identity (or lack thereof) of the same atom(s) in different substructures is known. For example, if there exist sets of substructures which are known not to overlap, then they can be specified as a single substructural constraint because GENOA never allows atom overlap within a given substructure. Another mechanism operates by "coloring" selected atoms with specified colors, where a color is simply another atom property like hydrogen range, hybridization and so forth. In this way atoms which are known to be distinct from other atoms can be differentiated from one another. The example given below illustrates both methods of controlling the degree of overlap.

#### *Diasin, a diterpene from Croton diasii*

Diasin is a rearranged labdanic diterpene derived from the trunkwood of *Croton diasii* Pires (Euphorbiaceae) [17]. This compound, obtained from the C<sub>6</sub>H<sub>6</sub> extract of the trunkwood, has the molecular composition C<sub>21</sub>H<sub>24</sub>O<sub>7</sub>. The following analysis of the spectral data to derive possible structures is similar to that employed by de Alvarenga et al. [17], in association with postulated biogenetic pathways, to derive the structure of diasin. The sub-

structural constraints inferred from a variety of spectral and chemical data are summarized in Fig. 2 as structural subunits 1--24. The results obtained by GENOA after specification of each constraint are summarized in Table 1 and described in detail below.

Infrared data revealed that the oxygens were incorporated into three ester groups and a furan ring. These substructural data were given to the GENOA program. The first constraint, 1 defines three distinct ester groups as a single substructure because it is presumed that they do not overlap. There is of course only one way to construct 1 (Table 1).

Analysis of the  $^1\text{H}$ -n.m.r. data showed the furan system to be  $\beta$ -substituted. The  $\beta$ -substituted furan system was defined as 2 and applied as the second constraint. There is only one way to construct 2 (Table 1).

One of the esters was in fact a methyl ester as evidenced by  $^1\text{H}$ -n.m.r. and  $^{13}\text{C}$ -n.m.r. data. Diasin includes only one other methyl group, this being attached to a quaternary alkyl ( $sp^3$ -hybridized) carbon. Constraint 3 defines the environments of the two methyl groups. Again there is only one way to realize these structural requirements (Table 1).

TABLE 1

Substructural constraints, specified range of occurrence and number of resulting cases for GENOA's analysis of the structure of diasin [17]

Substructure <sup>a</sup>	Range of occurrence	Resulting number of cases	Substructure <sup>a</sup>	Range of occurrence	Resulting number of cases
1	Exactly 1	1	12	Exactly 1	71
2	Exactly 1	1	13	Exactly 2	71
3	Exactly 1	1	14	Exactly 5	49
4	Exactly 1	4	15	Exactly 8	19
5	Exactly 1	308	16	None	17
6	Exactly 1	186	17	Exactly 1	14
7	None	166	18	None	14
8	Exactly 1	40	19	None	14
9	Exactly 1	65	20	At most 7	14
10	At most 1	65	21	At most 7	14
11	None	65			

Issue GENERATE Command → 145 Structures  
Transfer to STRCHK

		Resulting number of structures
22	At least 1	107
23	At least 1	5
24	At least 1	2

<sup>a</sup>See Fig. 2 for identity of substructure.

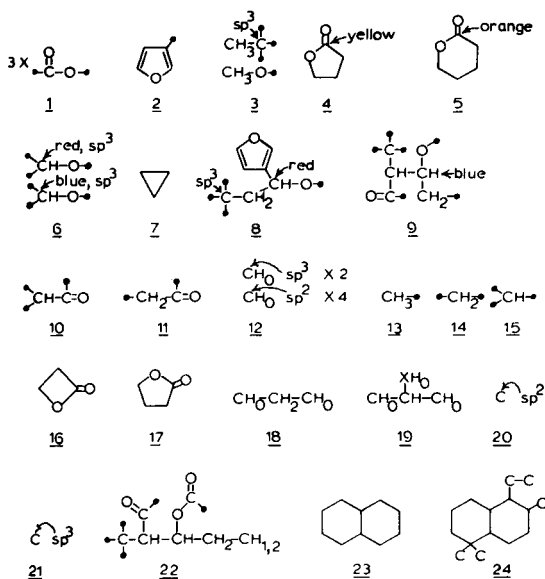
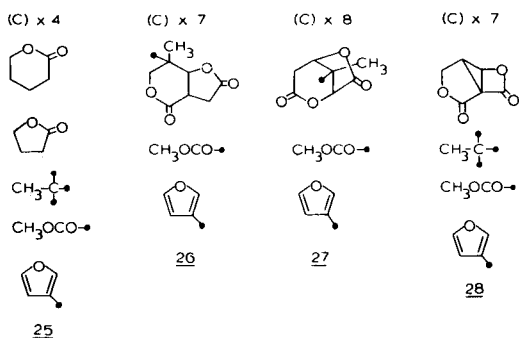


Fig. 2. Substructural constraints supplied to GENOA during establishment of the structure of diasin [17]. Bonds with an unspecified terminus (free valences) are to non-hydrogen atoms. If a substructure bears one or more free valences all other unfilled valences are assumed to be saturated with hydrogen atoms. If there are no free valences in a substructure, all unfilled valences may be connected to any atom including hydrogen.  $\text{CH}_0$  means a carbon bearing no hydrogen, e.g., in 18. The atom name "X", e.g., in 19, means any non-hydrogen atom.

Inspection of the i.r.-spectrum revealed that the other two ester carbonyls were incorporated in a  $\gamma$ -lactone system and a  $\delta$ -lactone system. Although the carbonyl atoms of these two lactone systems are definitely distinct, it is quite possible for the two lactone systems to overlap to some degree as would be the case if they were edge-fused. As mentioned previously, GENOA has a mechanism to handle such partial overlaps by assigning colors as atom properties to those atoms known to be distinct. In this case, the carbonyl carbons of the  $\gamma$ - and  $\delta$ -lactone systems were colored yellow and orange, respectively. In this way GENOA will never overlap the carbonyls but will allow the other atoms in the lactone ring (uncolored) to overlap, as illustrated below. Definition of the  $\gamma$ -lactone as 4 and use of it to construct new cases yields four results (Table 1) differing only in the selection of carbon atoms used to complete the five-membered ring. Definition and use of 5 yields 308 new cases (Table 1). These cases represent the large number of ways of realizing constraints 1–5 in the absence of additional data. Three cases are shown to illustrate the variety, 25–27. In 25 the lactone rings are completely distinct. In 26 they are edge-fused as indicated and in 27 the two lactone rings are realized via a bridged ring system.



Intermediate cases can be displayed at the discretion of the investigator at any point during an analysis. This is how 25—27 were obtained. Frequently, inspection of such structures will reveal the presence of undesired structural features that, although possible given the substructural constraints used, are in fact incompatible with known structural data. If such undesired features are observed then structures possessing them can be eliminated. This is illustrated later in this discussion in conjunction with constraint 7.

The next substructural constraint applied (6) specified the existence, inferred from  $^1\text{H}$ -n.m.r. and  $^{13}\text{C}$ -n.m.r., of two oxymethine atoms. Such atoms were distinguished by analysis of couplings in the  $^1\text{H}$ -n.m.r. spectrum. Therefore, since the methine carbons are distinct, constraint 6 was defined with different colors assigned to the carbons. This facilitates future reference to these particular methines in subsequent constraints where proton decoupling patterns relate to one or the other of the two distinct methine protons. Application of this constraint does eliminate some of the cases resulting from previous steps; the 308 cases are reduced to 186 (Table 1). Cases removed include those in which a lactone oxygen was attached to the quaternary carbon bearing the methyl group.

Here, inspection of intermediate results did reveal an unwanted structure. Some of the ways of combining the two lactone rings, e.g. as in 28, resulted in three-membered rings. The  $^1\text{H}$ -n.m.r. data for diasin are incompatible with a cyclopropyl group and so such structures are eliminated from further consideration by a new constraint (7) that prohibits rings of three carbons. Twenty cases of the original 186 are removed, yielding 166 cases (Table 1).

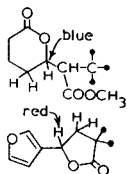
In the  $^1\text{H}$ -n.m.r. spectrum, the signal of one of the oxymethine groups was easily analyzed. This resonance was at characteristically low field ( $\delta$  5.45). Using shift reagents, de Alvarenga et al. [17] established that this oxymethine group was bonded to the furan ring and to a methylene group that showed no further coupling. Based on the n.m.r. data, this methylene must be bonded to a quaternary alkyl carbon. Substructure 8 expresses this constraint. The methylene group is colored red because it corresponds to the red methylene of substructure 6. Expressing 8 as a constraint yields 40 new cases (Table 1).



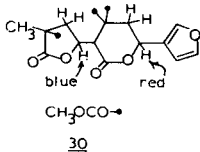
The signal multiplicity of the other oxymethine ( $\delta$  4.86) revealed its vicinity to three hydrogens: one ( $\delta$  3.1, d) on a carbon alpha to a carbonyl and a fully substituted position; and two represented by signals within a 12-proton envelope around 2 ppm. The appropriate substructure (9) expressing these inferences was defined, referring now to the blue methine of 6, and used as a constraint.

These constructive procedures resulted in rather more structural possibilities remaining at this stage of the analysis than were considered by de Alvarenga et al. [17]. The essential difference is that fewer assumptions have been made about the connectivity of the ester groups than were made by de Alvarenga et al. Thus, for example, GENOA continues to allow for the possibility that it is the  $\gamma$ -lactone that incorporates the oxymethine alpha to the furan ring, as expressed in case 29, in addition to the possibility considered by de Alvarenga et al. where the  $\delta$ -lactone is attached to the furan ring as expressed in case 30.

(C) x 2



(C) x 5

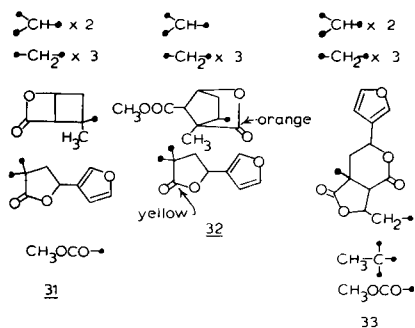


At this stage, most of the constructive inferences from the spectral data have already been made. But there still remain many additional inferences that can be made regarding the absence of particular substructures. Thus, there appears to be only one methine alpha to a carbonyl group and no methylenes alpha to carbonyls. Because the  $^1\text{H-n.m.r.}$  does not show any singlet resonances in the 12-H envelope (1.5–2.3 ppm) it can be assumed that there are no isolated methines or methylenes. Several constraints based on such inferences can be defined and applied. Since nothing has yet been stated regarding the bonding of the remaining methylene and methine groups, few of these additional constraints have any immediate effect. All will however be of value when it comes to constructing complete structures from the partially assembled cases. Constraints 10 and 11 express the absence of methines and methylenes alpha to carbonyl. As just mentioned, use of these constraints has no immediate effect on the number of cases (Table 1).

The rest of the information in the  $^{13}\text{C}$  spectrum, concerning the number of methyl, methylene and methine carbons, should also be used. Some of the cases constructed up to this point do in fact contain numbers of carbons whose degrees of substitution are in disagreement with the  $^{13}\text{C}$  spectra data. Constraint 12 expresses the number of quaternary carbons, two  $sp^3$  and four  $sp^2$  carbons. Constraints 13–15 express the observed numbers of methyl,

methylene and methine groups, respectively. Use of 12–15 as constraints yields eventually only 19 cases (Table 1).

Inspection of the partial structures obtained to this point revealed that in two the lactone systems had been bridged to give  $\beta$ -lactones, for example, 31. These structures were eliminated by constraint 16 (Table 1). (The GENOA command FORGET [16] would be more conveniently used here but use of the substructure constraint is given for completeness.) Similarly, in other structures the  $\delta$ -lactone ring system had been bridged giving rise to structures with two  $\gamma$ -lactone systems, one possessing the yellow carbonyl of constraint 4, the other possessing the orange carbonyl of 5, e.g., case 32. Definition and use of 17 eliminates such structures (Table 1). An example of one of the 14 remaining cases is 33, from which the structure proposed for diasin [17] was eventually obtained.



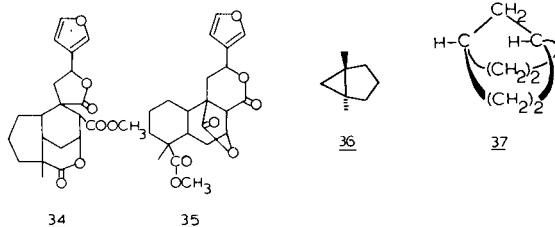
Constraints were then defined to prohibit the structure generation process from giving rise to final structures containing alkyl methines and/or methylenes that would give singlet resonances in the  $^1\text{H-n.m.r.}$  spectrum (18 and 19). Limits were also placed on the number of  $sp^3$ - and  $sp^2$ -hybridized carbon atoms (20 and 21). As before, these constraints have no immediate effect (Table 1) but will operate during generation of final structures to prevent undesired structures from being produced.

Finally, the structure generation process in GENOA was completed by issuing the GENERATE command. This process takes place under all of the constraints specified previously as GENOA checks to ensure that the specified range of occurrence of each substructure (1–21) is not violated. The total number of structures obtained is 145 (Table 1), reduced to 140 after elimination of structures possessing two  $\delta$ -lactones.

The GENOA program automatically transfers control to another component of the computer-assisted structure elucidation system. This second program, "STRCHK" is an updated version of the structure evaluation program described earlier [18]. It now includes various routines for inspecting results, checking for the presence of known skeletons, predicting spectra and analyzing stereochemistry. The 140 structures were checked against a small library of standard diterpene skeletons [10]. It proved that none of the generated structures incorporated any of these standard skeletons.

De Alvarenga et al. did not describe any further details of the coupling patterns of the protons. However, in their analysis they assumed that the methylene bonded to the oxymethine at 4.86 ppm (the blue methine of constraint 6) is not attached to a quaternary atom. Presumably, in the  $^1\text{H}$ -n.m.r. spectrum, this methylene showed some additional coupling. Therefore, an additional substructure, 22, was defined here in order to describe the environment more precisely. This constraint, 22, was applied to the set of 140 candidate structures using the PRUNE command [8], thereby reducing the set of structures remaining to be considered to 107 (Table 1).

De Alvarenga et al. made some additional assumptions concerning the molecular skeleton of diasin [17]. In particular, the assumption that the unknown is a rearranged labdanic diterpene would mean that the structure possesses at least a decalin ring system 23, if not the substituted decalin system 24. The structure checking program was used here to examine the remaining 107 structures to determine which, if any, incorporated either 23 or 24. Only five of the 107 incorporate the decalin system typical of diterpenes; two of these, 34 and 35, are based on the labdane skeleton represented by 24. The other three have the ring-A substituents shifted from the normal C-4 position.



Structure 35 corresponds to the structure proposed for diasin by de Alvarenga et al. [17]. It is probably not possible to eliminate structure 34 given just the spectral data for the unknown. Structure 34 may possibly be eliminated by biogenetic considerations. Further, 34 appears more sterically strained than 35 based on examination of molecular models. Certainly the changes in i.r. spectral data consequent upon epimerization of the unknown [17] are more readily rationalized in terms of 35.

Another approach to structure elucidation made possible with GENOA is one in which it is possible to assume that the unknown structure is based upon one of a set of possible skeletons. For example, the structure of diasin could have been studied beginning with a constraint like 24 to represent an initial assumption concerning the presence of a labdane skeleton. This type of problem occurs frequently in research on natural products when several structures have been characterized previously and it is possible to tell from preliminary physical and chemical data that a new structure is closely related to those already observed. Alternatively, GENOA can be initialized with a set of alternative substructures which represent several hypothetical skeletons on which a new structure might be based. Incorporation

tion of structural information through subsequent constraints will usually rapidly identify the correct skeletal type. If the new structure represents a more interesting problem involving a new skeletal type, then this is perceived at an earlier stage in GENOA as soon as additional constraints are found incompatible with the set of alternative skeletons.

## GENERATION OF CONFIGURATIONAL STEREOISOMERS

Simply knowing the constitution of a chemical structure is not in general enough to specify it for most purposes. Many properties of a molecule depend to some extent on its three-dimensional shape or stereochemistry. This problem is being studied by continuing the generation of isomers to include stereoisomers, in particular configurational stereoisomers (those which differ in configuration around chiral centers and double bonds). This section comprises brief descriptions of the exhaustive, irredundant generation of configurational stereoisomers, and the constrained generation of these stereoisomers; an example is given to illustrate the method.

### *Exhaustive generation of configurational stereoisomers*

The theory behind the specification and generation of configurational stereoisomers has been described in more detail elsewhere [19, 20], along with the computer program which implements it [21]. The first task is to find the atoms which are capable of being stereocenters. This is a difficult problem as there is no suitable definition of a stereocenter for these purposes. The method used bypasses this problem by rejecting all atoms which cannot be stereocenters and using the remaining atoms as candidates for stereocenters during the generation procedure. All appropriately substituted atoms and multiple bonds are processed in this way. In fact the atoms in multiple bonds are treated as individual stereocenters to provide an overall consistency to the approach [19, 21]. Configuration is defined using atom numbers which are provided by the generation of constitutional isomers described above. Thus, any nonplanar tri- and tetra-valent atom can be a stereocenter because all the numbers of the attached atoms differ; the problem of locating nonstereocenters is simply that of finding those atoms substituted in such a way that the two possible configurations are identical.

The problem of generation of configurational stereoisomers would now be simple if it were not for the problem of symmetry duplication. Each stereocenter can exist in two mirror image forms and for a structure with  $n$  stereocenters there would be  $2^n$  stereoisomers. However, many structures have symmetry and a significantly reduced number of distinct stereoisomers. This problem is solved by first finding the graph symmetry group of the structure. This is just the symmetry group of the constitutional isomer and is the entire group, not just the sets of equivalent atoms. The already located potential stereocenters are used to convert this group into a "configuration symmetry group" [19] which is the key data structure needed to generate, enumerate,

and specify the configurational stereoisomers. In addition, this group can be used to canonicalize substructures with configurational stereochemistry [22]. The distinct stereoisomers are generated using this group and a very efficient bit representation for the stereoisomers (each stereoisomer is effectively an integer in its computer representation). This symmetry group can also be used to simply enumerate (count without generation) the theoretical number of configurational stereoisomers using a novel enumeration formula [19]. This is done independently in the computer program to aid in specification of the scope of a structure elucidation problem and to provide an internal check on the correctness of the generation algorithm.

The computer program to accomplish the generation of stereoisomers was incorporated as part of the CONGEN and GENOA programs. A recent paper presents several examples illustrating the scope of stereoisomerism for selected molecular formulas and structures [21]. This exhaustive generator is now supplanted by the constrained generator described in the next section.

#### *Constrained generation of configurational stereoisomers*

For a computer program which generates isomers to be useful as an aid to structure elucidation, it must be capable of being constrained to generate only those isomers consistent with available data. Since there is in general an exponential (in  $n$ , the number of stereocenters) number of stereoisomers for any constitutional isomer, this number can get excessively large for typical structure elucidation problems. However, there are generally many constraints about the stereochemistry of an unknown structure, including chemical realizability, stereochemical information from n.m.r. and chiroptical methods. For these reasons, a constrained stereoisomer generation program is important. Recent developments of the STEREO program to accomplish this have been described in more detail [23].

The new STEREO program [23] allows description of three kinds of constraints which depend on stereochemistry. These are constraints which depend on: (1) atom type; (2) the (non)existence of substructures which have configurational stereochemistry designations; and (3) symmetry (chirality, equivalent atoms or substructures, etc.)

Atom type constraints are particularly easy to describe. Only certain types of atoms are capable of being stereocenters. In typical structure elucidation problems these are carbon atoms and certain substituted nitrogen atoms. The STEREO program recognizes these and allows the user to specify if non-carbon atoms in certain environments are capable of being stereocenters.

Substructure constraints are in general more useful but also more complicated. An investigator may know that certain substructures with stereochemistry exist with varying numbers of occurrences in the molecule. The program is capable of imposing this constraint on the generation and pruning of lists of stereoisomers.

The investigator may also know that an unknown structure is chiral (or achiral) and possesses several sets of equivalent atoms or substructures (or no

symmetry at all). The STEREO program is capable of imposing these constraints also.

Stereochemical constraints can be applied either to the generation of a list of stereoisomers or the pruning (elimination of some structures) of an already generated list. A number of these constraints can be applied prospectively (before generation) and are therefore used very efficiently. In such cases a stereoisomer is eliminated before it is generated. For example, many of the substructure constraints can be applied prospectively. Particularly useful constraints are those which eliminate highly strained stereoisomers. Structures with substructural features such as 36 and 37 are generally considered too strained to be realistic possibilities under most experimental conditions.

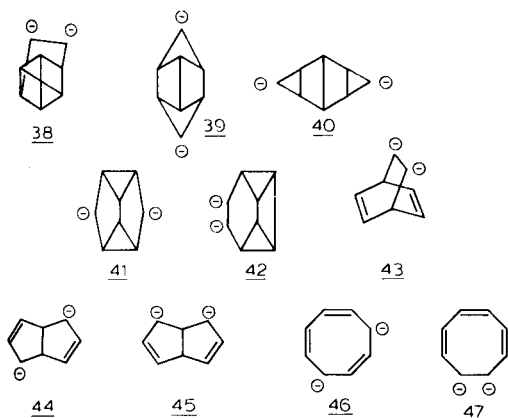
In these cases the substructure is found in the complete substructure by stereochemical graph-matching (matching constitution and configuration). During generation of stereoisomers the configurational constraints are applied to the prospective stereoisomer (represented as a bit-pattern or integer [19]) and any which do not pass here are not processed further. It is possible for constraints such as disallowing structures like 36 and 37 to eliminate all the stereoisomers of a single constitutional isomer [23] (example below).

To make use of constraints based on the symmetry of molecules, knowing the configuration of stereocenters is particularly important. In all cases the symmetry of a configurational stereoisomer will be less than or equal to the symmetry of a constitutional isomer. Prediction of properties (such as number of equivalent atoms or substructures) using only the constitutional symmetry is difficult since the constitutional symmetry is generally too large. Knowledge of the symmetry of the configurational stereoisomer is an improvement here but does not solve the problem of over-specified symmetry as the symmetry of a conformation can be still lower than the symmetry of a configurational stereoisomer. For this latter reason, the symmetry constraints in the STEREO program are generally "least only" constraints, i.e., the investigator specifies that there be at least  $n$  equivalent atoms of some type rather than exactly  $n$  equivalent atoms. The investigator can also specify that there be equivalent substructures (rather than just atoms). A subtle feature of this problem is that finding equivalent substructures requires the symmetry group of the configurational stereoisomer rather than just a list of equivalent atoms. Algorithms which only find sets of equivalent atoms would not be sufficient for this purpose.

*Example.* Recently the structural alternatives, including constitutional and stereochemical isomers, have been explored for a problem involving a  $C_8H_8$  dianion [24]. Constraints derived from physical and chemical data included the facts that each carbon, including the two negatively charged carbon atoms, bears exactly one hydrogen atom. There are no methyl groups or other substituents; all atoms appear to be included in a single ring system. Constitutional isomers were generated using CONGEN by defining a new atom type,  $C^-$ , with a valence of three, and using the modified molecular formula represented by  $(CH)_6(C^-H)_2$ . There are 46 structures which satisfy

the above constraints including a variety of ring systems possessing zero to three double bonds.

The 46 structures were then processed using the stereochemical constraints described below. Simply generating all the possible stereoisomers yielded 955 theoretical stereoisomers. This number is also obtained by simply enumerating them rather than generating them. Imposition of the constraint that there be no rings smaller than size eight with a *trans* double bond [23] reduces this to 46 constitutional structures with 690 stereoisomers. This can be done either by pruning the generated list or by constrained generation. Imposition of the constraint that there be no structures with 37 as a substructure reduces the total to 43 constitutional isomers with 208 stereoisomers. Note that three constitutional isomers have been eliminated entirely. An example of such a structure is 38.



It was observed in the  $^1\text{H}$ - and  $^{13}\text{C}$ -n.m.r. spectra of the unknown that there were three sets of equivalent atoms, one containing four atoms and two containing two atoms each. One interpretation of this observation is that there are four equivalent  $>\text{CH}-$  groups, a distinct set of two equivalent  $>\text{CH}-$  groups and two equivalent  $-\text{C}^-\text{H}-$  groups. Imposition of these constraints yields only five constitutional isomers with eight stereoisomers. These are 39–43 of which 40 has four possible stereoisomers and the others have one each. Note that 42 and 43 formally possess two adjacent carbanion sites and thus may be less likely candidates.

An alternative explanation of the equivalent sets of atoms observed in the n.m.r. spectra considers resonance structures for allyl carbanions. Under this assumption, the set of four equivalent atoms could be comprised of the four anion sites allowed by resonance, while the other two sets of two equivalent atoms are the two central atoms of the allyl carbanion and two other equivalent carbon atoms. This constraint can be simply expressed by requiring STEREO to retain only these stereoisomers which possess two equivalent substructures of the form  $-\text{C}^-\text{H}-\text{CH}=\text{CH}-$ .

Applying this constraint to the set of 43 structures (above) yields four

possible structures, 44–47 possessing, respectively, three, two, four and four stereoisomers. Structures 44 and 45 constitute an interconverting (by resonance) pair, as do 46 and 47, although, formally, 47 possesses two adjacent carbanion sites.

The information presented above was obtained to determine the scope of the problem and to guide new experiments to solve the structure. At this time, the problem has not been solved.

## SPECTRUM PREDICTION AND ANALYSIS

### *Mass spectra*

Of the various spectral techniques, mass spectrometry is generally the least influenced by stereochemical factors and thus the most amenable to analysis in terms of constitutional structure. Mass spectral/substructural correlations, such as those exploited in Heuristic DENDRAL's PRELIMINARY INFERENCE MAKER, are limited in their generality because they are tied closely to specific classes of chemical compounds. Structure  $\rightarrow$  spectrum relationships are somewhat more reliable and, as also demonstrated in the early Heuristic DENDRAL work, predicted mass spectra can usefully be employed to rank hypothesized candidate structures according to some measure of how well the predicted spectrum matches the observed spectrum.

A variety of methods for predicting mass spectra of given structures has been developed in parallel with the development of new structure generators. A simple model of molecular fragmentation, the "half-order theory," was initially developed as an aid to analyzing the characteristic fragmentations of sets of known structures [25]. The "half-order theory" generates all possible fragmentations of a structure compatible with constraints on the total number of bonds cleaved, the number of bonds cleaved in any single step, the number of cleavage steps and the permitted accompanying neutral losses and hydrogen transfers. Different plausibility values can be associated with each constraint and, thus, it is possible for the half-order theory to find all processes of at least some minimum plausibility that might give rise to an observed fragment ion in a given structure.

In addition to the general half-order theory, considerable effort has been expended on computer-aided methods for identifying specific cleavage processes characteristic of particular substructural features in a molecule [26]. Cleavage rules, identified by such programs, can allow fine discrimination between closely related isomers. The rules can identify how specific neutral losses and/or hydrogen transfers are induced by specific substructural forms and thus can capture much more subtle differences in fragmentation behavior than may be expressed in the half-order theory.

The use of the mass spectral prediction functions for structural analysis has undergone considerable development. The first implementation allowed structures created through CONGEN to be eliminated if, on the basis of a given theory for molecular fragmentation, they could not provide any



rationalization for individual chosen ions [12]. Subsequently, functions were developed that could estimate how well a given structure served to rationalize a complete observed mass spectrum. For each hypothesized structure, the mass spectral theories were applied to find the highest plausibility process which would lead to an observed ion from the structure. The score associated with a given candidate structure was based upon the sum of scores for individual ions with the individual scores based on the plausibility of the process and a measure of the importance of the ion in the spectrum. Typically, such scoring schemes separate candidate structures into two groups, those that provide approximately equally good rationalizations of the observed data and those that are unreasonable.

The various approaches to mass spectral analysis have now been combined into a system of programs that can employ a variety of predictive theories of differing degrees of specificity. Applications of these functions using half-order theory for predicting spectra of candidate structures have been illustrated for diverse compounds [27]. Applications using highly detailed, rule-based schemes for analysis of spectra for compounds in a known class have been illustrated using the spectra of marine sterols [28].

#### *Carbon-13 nuclear magnetic resonance spectroscopy*

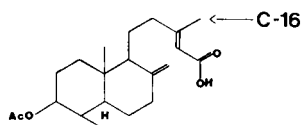
In many published reports on the elucidation of the structures of novel compounds, the analysis of  $^{13}\text{C}$ -n.m.r. data has been quite limited. Often, simply the number of methyl, methylene and methine groups from multiplicities and the number of unsaturated carbons from gross chemical shifts are reported. However, it is known that the chemical shift of a carbon nucleus is a sensitive probe of its stereochemical environment out to a three or four bond radius. Thus, there is considerably more structural information in a  $^{13}\text{C}$ -n.m.r. spectrum than is generally exploited.

The approach developed here for  $^{13}\text{C}$ -n.m.r. analysis exploits a data base of substructural environments and associated chemical shifts. The data base contains information, derived from known reference compounds, specifying the mean shift and shift range associated with carbon atoms in defined substructural environments. Similar data bases employing constitutional substructure representations have been developed previously [29, 30]. However, unlike mass spectrometry, carbon-13 spectral signatures are very sensitive to stereochemical influences. Thus, characterization of stereochemical substructure environments has been emphasized in developing our data base and associated programs for  $^{13}\text{C}$  spectral analysis. Recent developments of the STEREO program (see above) provided the tools for representing and manipulating configurational stereochemistry, extensions of these tools led to substructural characterizations incorporating such stereochemistry. Details of the procedures for defining stereochemical substructures and building and maintaining the data base have been presented [22].

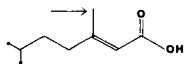
*Carbon-13 n. m. r. spectrum prediction.* The spectrum prediction procedures utilize the data base of substructures and chemical shifts as a kind of extended correlation table. The substructural environment of each carbon atom in a

candidate structure (e.g., from CONGEN, GENOA or defined manually, including stereochemistry) is described out to four bonds distance through the molecular skeleton, including configurational stereochemistry. This is done automatically, by using the same substructure coding scheme as for construction of the data base [22]. The substructure obtained is looked up in the data base and information is retrieved defining the range of shifts associated with atoms in similar substructural environments in previously analyzed reference compounds [31]. Thus, as illustrated in Fig. 3, the resonance shift from atom C-16 in the labdanoid structure shown is expected to be around 19.2–19.3 ppm. Note that the substructure possessing an  $\alpha$ ,  $\beta$ -unsaturated ester functionality as opposed to the corresponding acid functionality displays chemical shifts in the range 18.5–19.1 ppm (Fig. 3).

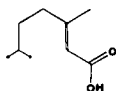
Of course, the four-bond substructural environment of an atom can include a large portion of a molecular structure. Consequently, it is common that when spectra are predicted for a particular candidate, some of the constituent atoms are found to be in substructural environments not represented in the current data base. Thus, for example, atom C-10 in the same labdanoid corresponds to a four-bond substructural environment new to the current data base. In such cases, the prediction processes try for more general three-, two- or even one-bond substructural environments [31]. As shown in Fig. 4 the best model that could be found for C-10 was in fact a two-bond substructural environment. The data in the data base characterizing these more



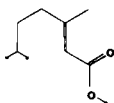
four-bond environment:



D  
A  
T  
A  
B  
A  
S  
E



19.2-19.3



18.5-19.1

$\delta_{\text{tms}}$

general substructures combine the shift ranges of all the more specific substructures that they subsume. Consequently, increasingly more general substructures are associated with increasingly wider shift ranges.

The result of this spectrum prediction process is a "fuzzy" spectrum in which atoms are associated with resonance (chemical shift) ranges rather than specific resonances [31]. The form of such fuzzy spectra is illustrated in Fig. 5 which shows the observed spectrum of pimaradiene (whose structure and spectrum are in the data base along with several closely related structures). The predicted spectrum using just two-bond substructural models displays resonance ranges varying from as little as 2.5 ppm to as much as 13 ppm, while predicted resonance ranges using four-bond substructural models are all less than 2 ppm.

To rank-order a set of candidate structures, such fuzzy spectra can be predicted for each member of the set. The structures can be represented as either constitutional isomers or stereoisomers. Functions have been developed that compute a score which is a reflection of the degree of compatibility between a predicted fuzzy spectrum and an observed spectrum [31]. These functions take into account the varying quality (four-bond, three-bond, ...) of the models used to predict the shift ranges. If an hypothesized structure yields poor agreement between its predicted and an observed spectrum, it is penalized heavily if the predictions were based on good models of its substructural environments (each atom characterized to a three- or four-bond level). Otherwise it is not penalized heavily because a structure cannot be

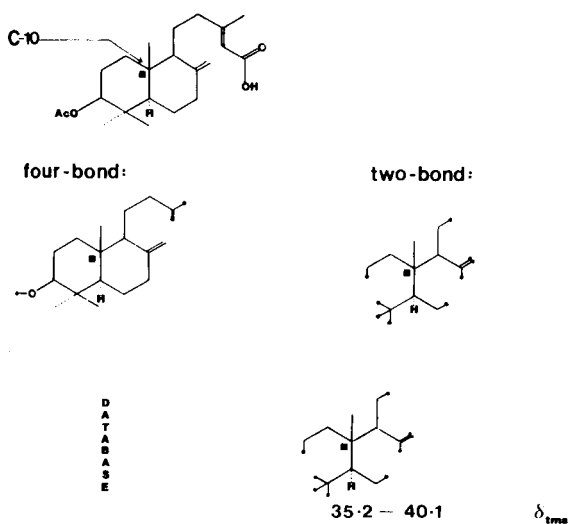


Fig. 4. Prediction of the resonance shift for atom C-10 of a labdanoid diterpene. The four-bond configurational stereochemical substructural environment is determined but found to be missing from the data base. Similar analysis for the three-bond environment also fails to yield a suitable prototype. A two-bond substructural environment is, however, present and associated with a resonance range of 35.2–40.1 ppm.

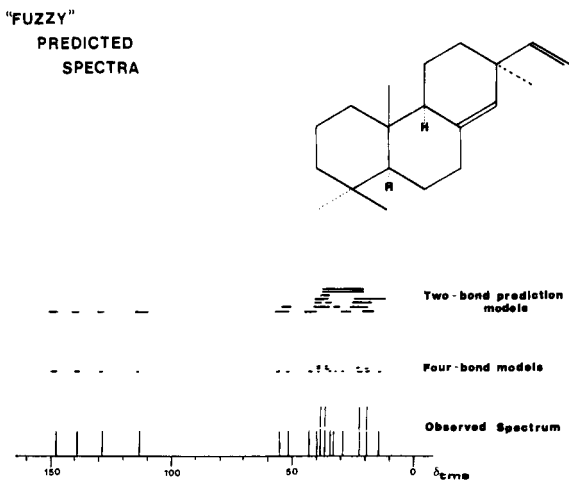


Fig. 5. A comparison of the observed spectrum and two "fuzzy" predicted spectra of pimaradiene, using both two-bond and four-bond substructural environments as bases for prediction of the resonance shift of each atom.

lowly ranked simply due to the inadequacy of the data base. The resulting scores can be used for structure ranking in order to differentiate between plausible and implausible structures.

*Carbon-13 n.m.r. spectrum interpretation.* The data base used for  $^{13}\text{C}$  spectrum prediction is also employed in the experimental approach to  $^{13}\text{C}$  spectrum interpretation. The computer-based interpretive procedures again treat the data base much like a very detailed correlation table [31]. However, there are usually many alternative substructures correlated with a resonance of given multiplicity and shift. Figure 6 shows a few of the different substructures that can be correlated with the occurrence of a quartet resonance in the range 14–15 ppm. No useful substructural constraints can be derived from such a wide variety of structural forms.

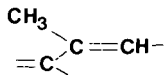
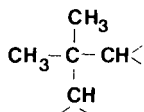
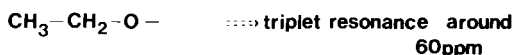
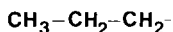


Fig. 6. Some of the alternative substructures found in the data base to be correlated with a quartet resonance in the range 14–15 ppm.

For any particular substructure to be a valid rationalization of one resonance from a complete  $^{13}\text{C}$  spectrum it must also be consistent with substructural interpretations found for other resonances [31]. For example, if a complete spectrum being analyzed did not show a triplet resonance in the range 50–70 ppm, then none of the methylene groups in that structure could be flanked by methyl and oxygen groups. Consequently, it would be possible to eliminate the second of the alternative substructures shown in Fig. 6 by comparing the substructures found for methyl resonances with those found for methylenes. This cross-checking and elimination procedure is automatic and iterative. The data base of substructures and shifts is searched to find possible interpretations for each individual resonance that are consistent with known bonding constraints. New bonding constraints are derived by intercomparison of the substructures retrieved [31]. These new constraints are applied in a subsequent iteration through the set of possible substructures. The data shown in Fig. 7 illustrate how, in one example [31], increasingly elaborate local substructural descriptions could be derived for one of the resonances of the complete spectrum being analyzed.

Figure 8 illustrates the form of the present  $^{13}\text{C}$ -n.m.r. interpretation program. The program takes as its primary input the molecular formula and complete  $^{13}\text{C}$ -n.m.r. spectrum of an unknown. The final output from the program consists of definitions of substructural constraints that can be utilized within GENOA. In addition to the basic  $^{13}\text{C}$ -n.m.r. data, an investigator can provide additional data. Such additional data can relate specific  $^{13}\text{C}$  resonances to known substructures, e.g., in the association of specified observed resonances at 21.3 and 171 ppm with an acetoxy group known to be present in the molecule (Fig. 8). It is also possible to employ more general constraints such as one specifying that the unknown structure must incorporate some standard skeleton, e.g., a labdane skeleton (Fig. 8).

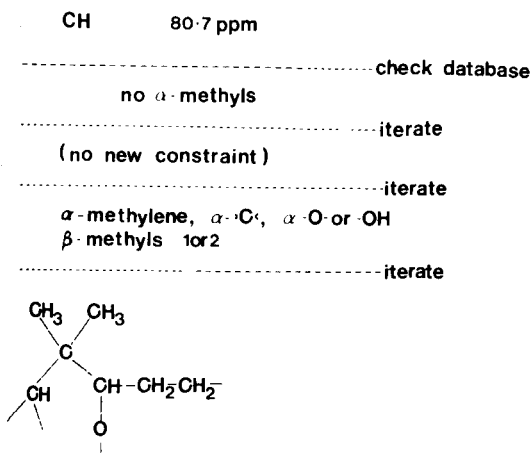


Fig. 7. Substructural environments for the  $>\text{CH}-$  resonance at 80.7 ppm, refined by application of the iterative cross-checking procedure taking into consideration the complete spectrum of a labdanoid diterpene [31].

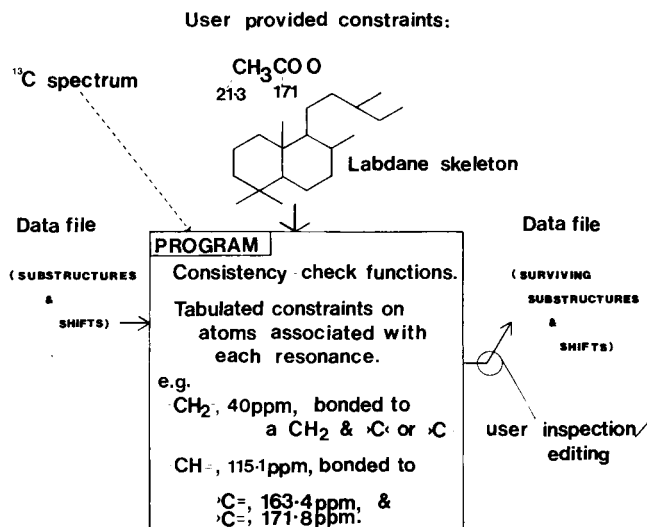


Fig. 8. Schematic form of the interpretive program for inferring substructural constraints from <sup>13</sup>C-n.m.r. data.

The program next accesses the data base to retrieve substructures consistent with each of the individual resonances. Substructure/shift combinations consistent with given (above) constraints are saved on a disk file. As each iterative step is completed, this file of surviving substructures and shifts becomes the input for the next iteration. Constraints derived automatically by the program during these iterations may specify just the atom type of allowed neighbors or may define specific bonding patterns. In this way, substructural environments tend to "grow" larger by one or more atoms as the program progresses. Because the program is interactive, an investigator can contribute to the iterative analysis by inspecting the substructures that the program is considering for each individual resonance, and eliminating any that are inappropriate on the basis of other data.

A typical application of this program is illustrated by the analysis of the <sup>13</sup>C-n.m.r. spectrum shown in Fig. 9 under the constraint, derived from other chemical and physical data, that the unknown is a substituted cholestane [32]. In this example, the analysis that the <sup>13</sup>C-n.m.r. interpretation program must achieve is essentially that of deriving constraints limiting the placement of various substituents upon the given cholestane skeleton. The information derived at various stages in the iterative analysis is summarized in Fig. 10.

The initial state, represented by *A* in Fig. 10, is inferred from the constraints of the molecular formula and the number of observed methyl resonances. In *A*, '?' identifies a carbon atom whose degree of substitution is still to be determined. The first check of the observed spectrum against a data base of sterols and analysis of the derived constraints yielded the more completely elaborated structure *B*. Further iterations yielded first *C*, in

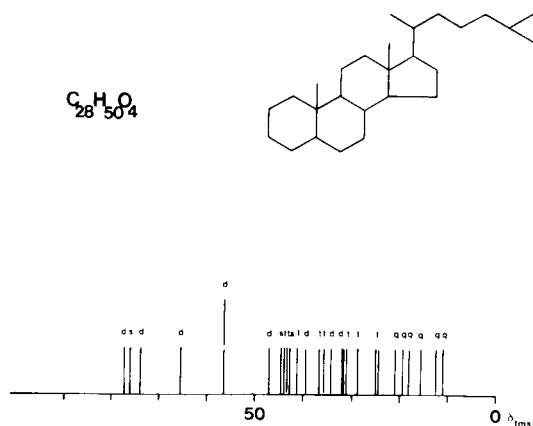


Fig. 9.  $^{13}\text{C}$ -n.m.r. spectrum of a novel marine sterol, a polyhydroxy cholestane derivative [32].

which the program determined that all the hydroxy groups must be on the A or B rings, thus placing the remaining methyl at C-24, and finally structure *D*. Structure *D* in fact represents the constitution of this sterol. The additional methyl group is at C-24 while the four remaining hydroxyl groups are at C-1, C-3, C-5 and C-6. The program has been able to associate specific observed resonances with every atom marked with asterisks in *D*, thus effectively performing a partial assignment of the spectrum as a by-product of the interpretation.

In this sterol example, the combination of initial data which pointed toward the cholestane skeleton and the  $^{13}\text{C}$ -n.m.r. spectrum was sufficient to establish unambiguously the constitution of the structure. More typically, available data are less precise but still sufficient to allow the  $^{13}\text{C}$ -n.m.r. interpretation program to derive descriptions of the allowed substructural environments of some or all of the constituent carbon atoms. These substructures can then be used, in association with any other substructural constraints

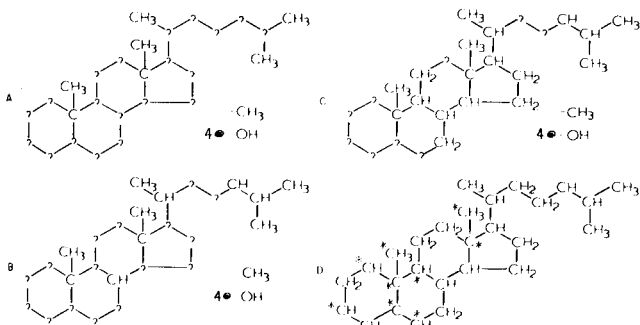


Fig. 10. Data inferred about substitution of the cholestane skeleton at different stages of the iterative analysis.

available from other sources, in the GENOA isomer generation program. It is characteristic of the  $^{13}\text{C}$ -n.m.r. spectrum interpretations that the resulting substructures overlap considerably. Thus, GENOA can be used directly to generate structural candidates while CONGEN is much less useful because considerable manual effort would be required to derive superatoms which did not overlap one another.

*Limitations.* Although potentially of considerable value, the  $^{13}\text{C}$ -n.m.r. analysis functions have certain intrinsic limitations. The entire approach does depend on having a complete  $^{13}\text{C}$ -n.m.r. spectrum with an observed resonance of known multiplicity for every constituent carbon atom. Frequently however, the amount of a novel natural product available for analysis is insufficient to allow such data to be acquired.

The  $^{13}\text{C}$ -n.m.r. prediction and structure-ranking procedures are appropriate when at most two to three hundred candidate isomers must be processed. These procedures can only usefully discriminate between different candidates in cases where the data base contains substructures representative of structural features of the isomers wherein there are structural differences.

The experimental interpretation functions are even more sensitive to limitations in the data base. The interpretive analysis yields useful results only when the data base is sufficiently broad to have representative substructures for observed resonances. Generally, to be a useful procedure for limiting the total number of candidates, there must be an appropriate substructural model for each resonance matching to at least a two-bond radius. Although this does not sound a severe constraint, for it corresponds to considering just  $\alpha$ - and  $\beta$ -influences on shifts, given the vast variety of structural forms possible it is common for new structures to contain substructural features that are distinct at a two-bond radius from all those currently registered in the data base. The program will report failure if it finds inconsistencies resulting from attempts at interpretation that assume that the data base contains better substructural models than are in fact present. In any case, structures derived from such interpretive procedures followed by structure generation in GENOA are only hypothetical. They still have to be proven by conventional approaches such as X-ray crystallography, unambiguous synthesis or conversion to a previously characterized material.

## DISCUSSION

In GENOA, the structural chemist has for the first time a computerized structure generation procedure that can utilize structural information, as it is acquired, without any pre-requisite for detailed interpretation by the chemist. The flexibility of the structural constraints allowed in GENOA, coupled with the ability to test simultaneously many different structural hypotheses, should make GENOA a powerful tool at all stages in structure analysis. The power of programs such as GENOA and CONGEN is now further enhanced through the STEREO program, which permits the complete



analysis of the potential configurational stereochemistry of a constitutional isomer.

Methods for testing and ranking generated structures appear now to be fairly satisfactory. With mass spectra, the theories used for spectrum prediction can vary from the very general and widely applicable "half-order" theory to highly specific rule-based systems for structures of a particular class. These methods have been widely applied and suitable ranking functions have been developed for the different mass spectrum prediction functions. The prediction/ranking functions for  $^{13}\text{C}$ -n.m.r. are less highly developed. In particular, the ranking functions require further development in the area of generally applicable scoring schemes for comparing predicted and observed spectra. In addition, the  $^{13}\text{C}$ -n.m.r. procedure has nothing quite comparable to the almost universally applicable "half-order" theory of mass spectrometry and will always be somewhat limited by the quality of the available data base of substructures and shifts.

Methods for inferring substructural constraints from spectral data are still limited in scope. In addition to the  $^{13}\text{C}$ -n.m.r. systems outlined above, some limited studies on exploiting mass spectral data have also been undertaken [33]. Like the  $^{13}\text{C}$ -n.m.r. system, these mass spectral "interpretation" functions rely on a data base correlating spectral features with particular substructures. The most suitable mass spectral data for use in this approach are those obtained from m.i.k.e.s. [34] or m.s./m.s. [35] of the individual ions resulting from the fragmentation of a molecule. Construction of a data base of m.i.k.e. spectra for many different fragment ions from many classes of compounds would be time-consuming. Recently, however, a proposal to build and maintain such a data base has been presented [36].

There are a number of on-going developments arising out of work presented here. In current programs, configurational stereochemistry is relegated to a subsidiary role (following the generation of constitutional isomers). Although this parallels some problems in structure elucidation, where first the molecular constitution is determined, followed by investigation of stereochemistry, this approach makes it difficult to utilize stereochemical constraints early in the programs. The fact that stereochemical constraints may reject many constitutional isomers in addition to stereoisomers has been previously discussed [23]. In future systems, a more consistent approach will unify the handling of configuration and constitution.

Another intended development is to extend these techniques to treatment of conformational stereochemistry. The various steps required in our approach to characterizing the three-dimensional aspects of chemical structure may be outlined as follows. The structure generation procedures are now capable of beginning with the molecular formula and, using CONGEN or GENOA, determining the complete set of constitutional isomers under constraints. Subsequently, the complete set of configurational stereoisomers can be determined under constraints by using the STEREO program. As yet, there are no means available for developing conformational descriptions

from configurational data. There are, however, extensive existing methods, developed by many other groups, which are capable of converting descriptions of molecular conformations into three-dimensional coordinates. Using these coordinates a variety of energy calculations can be performed to refine further the structural descriptions. We intend to fill in the missing link between configuration and conformation in the sequence of procedures by developing general programs which can generate possible conformations of a structure(s) given its constitution and configuration. Such extensions will allow spectral data to be analyzed where spectral signatures are an intimate function of molecular conformation. In particular, conformational sub-structural descriptions may permit  $^1\text{H-n.m.r.}$  data to be analyzed by methods similar to those currently employed for  $^{13}\text{C-n.m.r.}$  data. In addition, proper representation of conformation will open the way for many other studies relating molecular structures to properties or activities other than spectral signatures.

## REFERENCES

- 1 J. Lederberg, DENDRAL-64, Part I. Notational Algorithm for Tree Structures, NASA Star No. N65-13158, NASA CR-57029 (1965); Part II. Topology of Cyclic Graphs, NASA Star No. N66-14074, NASA CR-68898 (1966); Part III. Complete Chemical Graphs: Embedding Rings in Trees, NASA Star No. N71-76061, NASA CR-123176 (1971).
- 2 J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield and C. Djerassi, *J. Am. Chem. Soc.*, 91 (1969) 2973.
- 3 B. G. Buchanan, G. L. Sutherland and E. A. Feigenbaum, in B. Meltzer and D. Michie (Eds.), *Machine Intelligence*, Vol. 4, Edinburgh University Press, Edinburgh, 1969, p. 209.
- 4 A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum and J. Lederberg, *J. Am. Chem. Soc.*, 91 (1969) 2977.
- 5 G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum and J. Lederberg, *J. Am. Chem. Soc.*, 91 (1969) 7440.
- 6 J. Lederberg, in G. R. Waller (Ed.), *Biochemical Applications of Mass Spectrometry*, Wiley, New York, 1972, p. 193.
- 7 L. M. Masinter, N. S. Sridharan, J. Lederberg and D. H. Smith, *J. Am. Chem. Soc.*, 96 (1974) 7702.
- 8 R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, *J. Am. Chem. Soc.*, 97 (1975) 5755.
- 9 See, e.g.: (a) C. J. Cheer, D. H. Smith, C. Djerassi, B. Tursch, J. C. Braekman and D. Daloz, *Tetrahedron*, 32 (1976) 1807.  
(b) M. J. Goldstein, Y. Nomura, Y. Takeuchi and S. Tomoda, *J. Am. Chem. Soc.*, 100 (1978) 4899.  
(c) D. J. Vanderah, N. Rutledge, F. J. Schmitz and L. S. Ciereszko, *J. Org. Chem.*, 43 (1978) 1614.  
(d) D. E. Dorman, Abstracts of the 178th National Meeting of the American Chemical Society, Washington, DC, Sept. 9-14, 1979, Computers in Chemistry Division, Papers 52 and 58.
- 10 D. H. Smith and R. E. Carhart, *Tetrahedron*, 32 (1976) 2513.
- 11 R. E. Carhart, T. H. Varkony and D. H. Smith, in D. H. Smith (Ed.), *Computer-Assisted Structure Elucidation*, ACS Symposium Series 54, American Chemical Society, Washington, D.C., 1977, p. 126.

- 12 D. H. Smith and R. E. Carhart, in M. L. Gross (Ed.), *High-Performance Mass Spectrometry: Chemical Applications*, ACS Symposium Series 70, American Chemical Society, Washington, DC, 1978, p. 325.
- 13 C. A. Shelley, H. B. Woodruff, C. R. Snelling and M. E. Munk, in D. H. Smith (Ed.), *Computer-Assisted Structure Elucidation*, ACS Symposium Series 54, American Chemical Society, Washington, DC, 1977, p. 92.
- 14 T. Yamasaki, H. Abe, Y. Kudo and S. Sasaki, in D. H. Smith (Ed.), *Computer Assisted Structure Elucidation*, ACS Symposium Series 54, American Chemical Society, Washington, DC, 1977, p. 108.
- 15 L. A. Gribov, M. E. Elyashberg and V. V. Serov, *Anal. Chim. Acta*, 95 (1977) 75.
- 16 R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse and C. Djerassi, *J. Org. Chem.*, 46 (1981) 1708.
- 17 M. A. de Alvarenga, H. E. Gottlieb, O. R. Gottlieb, M. T. Magalhaes and V. O. DaSilva, *Phytochemistry*, 17 (1978) 1773.
- 18 C. Djerassi, D. H. Smith and T. Varkony, *Naturwissenschaften*, 66 (1979) 9.
- 19 J. G. Nourse, *J. Am. Chem. Soc.*, 101 (1979) 1210.
- 20 J. G. Nourse, in J. Hinze (Ed.), *The Permutation Group in Physics and Chemistry*. Springer-Verlag, New York, 1979, p. 19.
- 21 J. G. Nourse, R. E. Carhart, D. H. Smith and C. Djerassi, *J. Am. Chem. Soc.*, 101 (1979) 1216.
- 22 N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith and C. Djerassi, *Org. Magn. Reson.*, 15 (1981) 375.
- 23 J. G. Nourse, D. H. Smith, R. E. Carhart and C. Djerassi, *J. Am. Chem. Soc.*, 102 (1980) 6289.
- 24 M. J. Goldstein, private communication.
- 25 D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, *Tetrahedron*, 29 (1973) 3117.
- 26 B. G. Buchanan, D. H. Smith, W. C. White, R. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, *J. Am. Chem. Soc.*, 96 (1976) 6168.
- 27 N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. H. Varkony, B. G. Buchanan, W. C. White and L. Creary, *Anal. Chem.*, 52 (1980) 1095.
- 28 A. Lavanchy, T. H. Varkony, D. H. Smith, N. A. B. Gray, W. C. White, R. E. Carhart, B. G. Buchanan and C. Djerassi, *Org. Mass Spectrom.*, 15 (1980) 355.
- 29 (a) W. Bremser, M. Klier and E. Meyer, *Org. Magn. Reson.*, 7 (1975) 97.  
(b) W. Bremser, L. Ernst and B. Franke, *Carbon-13 NMR Spectral Data*, Verlag Chemie, Weinheim, Germany, 1978.
- 30 B. A. Jezl and D. L. Dalrymple, *Anal. Chem.*, 47 (1975) 203.
- 31 N. A. B. Gray, C. W. Crandell, J. G. Nourse, D. H. Smith, M. L. Dageforde and C. Djerassi, *J. Org. Chem.*, 46 (1981) 703.
- 32 Y. Yamada, S. Suzuki, K. Iguchi, H. Kikuchi, Y. Tsukitani, H. Horiai and H. Hakanishi, *Chem. Pharm. Bull.*, 28 (1980) 473.
- 33 N. A. B. Gray, A. Buchs, D. H. Smith and C. Djerassi, *Helv. Chim. Acta*, 64 (1981) 458.
- 34 F. W. McLafferty, R. Kornfeld, W. F. Haddon, K. Levsen, I. Sakai, P. F. Bente III, S.-C. Tsai and H. D. R. Schuddemage, *J. Am. Chem. Soc.*, 95 (1973) 3886.
- 35 R. A. Yost and C. G. Enke, *Anal. Chem.*, 51 (1979) 1251A.
- 36 F. W. McLafferty, A. Hirota and M. P. Barbalas, *Org. Mass Spectrom.*, 15 (1980) 327.

## COMPUTER-AIDED STRUCTURE ELUCIDATION METHODS

H. ABE, T. YAMASAKI\*\*, I. FUJIWARA and S. SASAKI\*

*School of Materials Science, Toyohashi University of Technology, Tempaku, Toyohashi, Aichi, 440 (Japan)*

(Received 23rd January 1981)

### SUMMARY

Computer-aided structure elucidation methodology is discussed. Major processes involved in computer-aided structure elucidation systems are partial-structure elucidation, structure generation, and structure examination. For the three representative systems CONGEN, CASE, and CHEMICS, these processes are examined. There are four necessary conditions for automated chemical structure elucidation systems: reliability, width of application, ease of modification and portability.

It is a fairly long time since computers came to be used as a tool for information processing. Studies on information processing are also actively carried on in various fields of chemistry [1]. Computers are now regarded as indispensable for calculating various physical quantities related to molecular electronic properties, chemical reactions or the like. Of course, processing of information began almost simultaneously with the development of "science". Methods of information processing are, from the viewpoint of data handling, classified into five specialized items as shown in Table 1. Those items are normally studied in connection with each other.

A general view may be taken as follows, wherein problems are limited to chemical information. Representation of information is regarded as the basis of the other items and the major problem to be considered is how to represent data that are of multidimensional structure. A chemical structural formula is a typical example of multidimensional data. Several methods based on linear notation [2], connectivity matrix, and graph theory have been proposed for representation of structural formulae [3], wherein the chemical structures are replaced by topological relations for the representation. Analog data are not suitable for computer processing, and must be converted to digital values. Major problems in acquisition of the data are analog-to-digital conversion of various spectral information and improvement of signal-to-noise ratio. The data thus obtained are then processed with empirical rules and/or theories. Calculation processes based on quantum theory are almost inconceivable without a computer. Handling of the processed result in this

---

\*\*Present address: Mitsui Petrochemical Co. Ltd., Iwakuni, Yamaguchi 740, Japan.

TABLE 1

## Handling of chemical information

Handling	Information
Representation	Structural formula, chemical reaction
Acquisition	Analog data via spectrometer
Processing	Quantum chemical data, spectrometric data, structural data
Display	Structure, chemical equations, graphical data
Store and retrieval	Bibliographies of chemistry

case is essentially the reverse of the acquisition of the data, and the results are displayed multidimensionally by inverting internal linear data inside the computer. Display of the three-dimensional structure of protein on a cathode-ray tube is a good example of this case [4].

In contrast, data have to be stored in extremely concentrated forms for convenience in information retrieval [2]. Elucidation of organic structures and estimation of various physical quantities concerning organic reactions have been studied as problems to be handled by integrating above five items [5]. These problems involve consolidated information processing, and solution of the problem or completion of the program system provides a breakthrough for establishing a methodology on the handling of the chemical information. In the early stages, the computer was expected to have functions similar to those of the human brain [6]. However, computers cannot achieve the flexible thinking faculty of the brain with free use of memorized information as background.

Information is generally processed along the flow lines shown in Fig. 1. Here, use of the computer is restricted to the process shown as Step 4, Processing information according to the pre-defined algorithm. The processing capacity of this is estimated by the logicity or consistency of the process conforming to a given formula (characteristic 1 of computers), by the processing speed (characteristic 2), and by the quantity processed (characteristic 3). It is quite easy to show that the function of computers is superior to the ability of the human brain in any of these three factors. For instance, characteristic 1 of the computer may be regarded as a weak point of the human brain. As for the other two characteristics, the differences are almost self-evident. Accordingly, the role of computers lies in applying completely and quickly given logical operations to a large quantity of information.

Formulation of the processing procedure or establishment of algorithms is the most important problem of the flow diagram shown in Fig. 1. So far as human thinking is logical, the procedure ought to be resolved into elementary processes, and when the problem is simplified in this way, computerization of human thinking should be possible. The approach to a problem is considered to comprise, first, the recognition of elementary processes through analysis of the problem, and then formulation of the elementary processes. Application

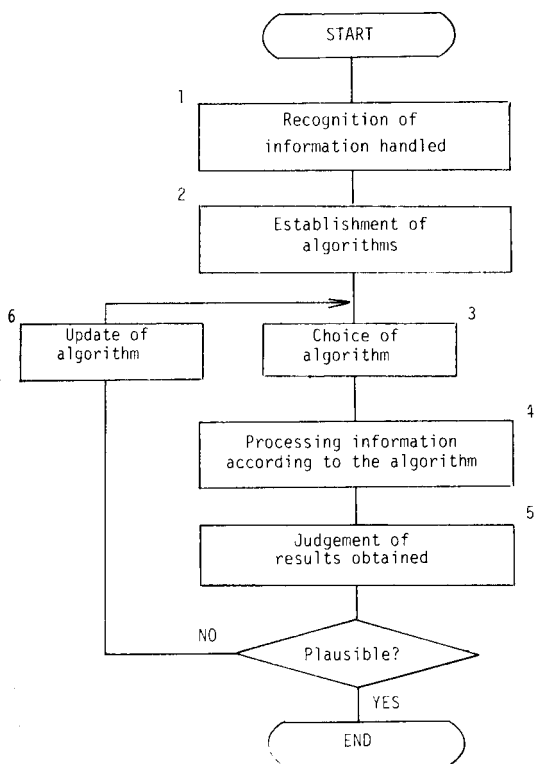


Fig. 1. Flow chart of handling of information.

of Fig. 1 to an information-processing problem involving elucidation of organic chemical structures is described below.

### STRUCTURE ELUCIDATION

Elucidation of organic structures is regarded as a system composed, in a complicated manner, of information-analyzing processes and processes totalizing the knowledge extracted; computerization of the system is therefore a challenging problem. For computerization, the given problems are resolved into elementary processes independent of one another initially. The five basic processes in structure elucidation may be regarded as analogous to the steps shown in Fig. 1, requiring an input unit, a partial-structure elucidation unit, a structure generation unit, a structure examination unit, and an output unit. According to the methods of data handling described above, the input unit, the output unit, and the other units pertain to data acquisition, data display, and data processing, respectively. The techniques used for these basic processes or data processing modes are many and varied.

Various kinds of data processing are listed in Table 2. Spectroscopic data are regarded as the data providing information on linkages of the partial structures or functional groups, and most of the data processing is concerned with analysis of these spectral data as shown in the table. Once the basic processes and the substance of the data processing in structure elucidation have been thus established, each process can be represented with an integrated algorithm developed for processing individual data. These processes can be executed by computers. The structure elucidation process is carried out flexibly by a chemist utilizing specific features of the various kinds of techniques aforementioned and applying such operations as feedback to the five processes mentioned above. The quality of the management used in dealing with such situations on the basis of intuition is regarded as one of the main differences between skilled and unskilled persons in structure elucidation. Such flexible action, by its nature, cannot be expected from computers; accordingly, a system (man-machine system) for compensating the mutual deficiencies of both man and machine should be thought out. Thus an ideal Information Processing Based on Algorithm will be realized by combining the flexibility or generality of human thinking with the massive and rapid processing ability of computers. Several methods for automation of structure elucidation, especially the individual data processing that forms the major portion thereof, are reviewed below.

#### *Methods for automation*

Depending on the kinds of data, data processing can be roughly divided into two categories, i.e., for spectral information and for structural information. The former is mainly used in partial structure elucidation and the

TABLE 2

Relation between process and data processing<sup>a</sup>

Processing	Process		
	Elucidation of substructure	Generation of structure	Examination of structure
Translation of spectral data	○	△	○
Translation of chemical reaction	○	△	○
Pattern recognition	○	—	—
File retrieval	○	△ <sup>b</sup>	○
Prediction of spectrum	—	—	○
Exhaustive enumeration of chemical graph	—	○	—

<sup>a</sup>(○) Direct application for each process; (△) partial application. <sup>b</sup>Exact meaning is not generation but identification of structure.

structure examination process, while the latter is used in the structure generation process. Examination of the spectral information is further subdivided into three items: matching, interpretation, and prediction of spectra. Matching is a technique used mainly for information retrieval, where input data are matched to previously stored large quantities of data (data base) and structures are identified by establishing similarities. There are many kinds of spectral data base for matching, such as mass spectra [7], i.r. spectra [8] and n.m.r. spectra [9]. Information retrieval systems have been developed [10, 11] where different kinds of spectra are used to complement each other, and some original methods have been applied in data storage procedures, in feature extraction from spectra, and in systems for estimating similarity, so that data processing in "chemist's language" has become possible.

Prediction of spectral data needs known structures as input data and is applicable to m.s., i.r., n.m.r. and u.v. spectra. Mass spectral prediction by means of empirical rules has been described [12]. Wilson's method [13] and the PPP (Pariser—Parr—Pople) method [14] are applicable to i.r. and u.v. spectrum, respectively. Prediction of  $^1\text{H}$ -n.m.r. spectra by the semi-empirical BBC (Bothner-By, Castellano) technique and by the Lindeman—Adams method [15] is possible. Both techniques are capable of producing relatively precise results but they are not widely applicable. Chemical shift data is fundamental for interpretation of n.m.r. spectra. Spin—spin coupling provides valuable information for  $^{13}\text{C}$ -n.m.r. spectra. In  $^1\text{H}$ -n.m.r.,  $A_m X_n$  patterns were analyzed by Beech et al. [16].

Interpretation of i.r. spectra is not yet completely formulated. For the interpretation of complicatedly superimposed absorption bands of i.r. spectra, Gribov et al. [17] developed a method in which partial structures are correlated with absorption bands and logically matched partial structures are designated by symbolic logic. Munk, Gray and others tried to elucidate structures with a previously prepared correlation table of partial structures and absorption bands [18, 19].

In the case of mass spectra, automation of analysis was attempted at a relatively early stage [20] because the data are essentially digital. Buchs et al. [21] developed algorithms for elucidation of ketones, ethers, and amines; simple empirical rules for electron impact-induced cleavage of organic compounds were mainly formulated. There are several studies in which pattern recognition methods are used for structure elucidation. Partial structures can be obtained, at an accuracy of 80% or better by using pattern recognition [22].

Structure generation is fundamental in processing structural information. In structure generation, all the probable structures are made up without any overlapping or omission of structures on the basis of given partial-structural information. Three kinds of generating system have been described, each of which has its own way of representing the structure. Nelson et al. [23] used a pair set composed of uncoupled bonds of the partial structure. Kudo and



Sasaki [24] used a connectivity stack, and Masinter et al. [25] used graph theory for representing structures. Basic algorithms and their main features are shown in Table 3. In the program CMBN, some partial structures are represented by serial numbers of semi-bondings, and full structures are enumerated through all the combinations of the semi-bonding couples. In contrast, the generation process in CONGEN starts with distribution of nodes to cyclic and acyclic portions of structures on the basis of the molecular formula, and then, in due order, rings are formed, chain structures are formed, and finally rings are coupled with chains. This method is based on graph theory, and the ease of logical verification is its distinctive feature.

#### SYSTEM FOR STRUCTURE ELUCIDATION

The basic processes of structure elucidation, the data to be used, and their processing, have been outlined above. The total structure elucidation system is built from those elements. Structure generation is the most difficult process for automation in structure elucidation, and the ability to eliminate overlapping or omission is the key to successful development of such a system. A computer-aided structure elucidation system should also meet certain conditions with respect to reliability, breadth of application, ease of modification, and portability from one computer to another.

With regard to reliability, it is desirable that the correct answer should be given for correct input data without requiring additional professional (chemical) knowledge from the user. If the system involves an interactive mode of operation so that various forms of structural knowledge can be input, consistency between those items of information should be checked automatically. Regardless of causes, erroneous answers reduce confidence in the system because users expect chemically reasonable answers rather than logically correct answers which are results of wrong operations. The chance

TABLE 3

Algorithm of structure generation

	Basic algorithm	Check for duplication	Node	Edge
CMBN (M. E. Munk)	Combination of "half-bond" pair	Topological symmetry and naming algorithm	Arbitrary	Arbitrary
BLD (S. Sasaki)	Exhaustive enumeration of canonical connectivity matrices	Permutation of connectivity matrix	Arbitrary	Mono-chromatic
CONGEN (D. H. Smith)	Assignment to vertex graph according to property of node (in the case of ring formation)	Reduction to coloring problem	Arbitrary	Mono-chromatic

of wrong operation increases as the amount of information required from the users increases.

The next important condition is breadth of application. How many kinds and numbers of elements are allowed in objective molecules determines the applicability of the system. In general, as the kinds and numbers of elements in objective molecules increase, the amount of responding answers also becomes larger. Therefore, there should be some restrictions, e.g., to compounds containing only carbon, hydrogen, and oxygen, or only carbon, hydrogen, nitrogen and oxygen. These limitations should not, however, be so severe that users feel the system to be useless in practical situations.

Because the algorithms for automatic interpretation of spectral data depend on empirical rules, occasional modifications are required to follow progress in the spectrometries used. Therefore it is desirable that the system consist of a set of highly modularized program units so that modification of a particular unit will not affect the others.

Finally, ease of conversion to other computer systems is indispensable for any scientific programs because as long as they are the result of research work, reconfirmation by other research workers is essential. For that reason, programs written in special languages and/or depending on special machines are not desirable.

Successful results have been achieved by systems such as CASE (Munk and co-workers [26]), CONGEN (Smith and co-workers [27]) and CHEMICS (present authors [28]). Algorithms are provided for all the major processes: partial structure elucidation, structure generation, and structure examination in CASE, and structure generation and examination in CONGEN. However, it is left for users to combine those processes functionally. Therefore, for the two systems, reliability depends on how efficiently the users can set up a group of partial structures without overlapping. CHEMICS is the only system in which not only the individual process, but even the information transfer between processes is included in the algorithms, and the users can obtain the structural formula merely by putting spectral data into the input unit. In other words, CHEMICS uses basically a batch processing method. At present, the system is reorganized for the interactive mode and allows users to input various partial structures obtained from other sources. However, the partial structures input by the users are checked automatically to establish if they are consistent with previously input spectral data.

Concerning its applicability, CHEMICS yields to CASE and CONGEN because it can handle only those compounds which contain C, H and O. With regard to ease of modification, we have insufficient information about CASE, and CONGEN is not provided with an automatic interpreter so that it is exempt from this discussion. CHEMICS is not yet ready for modification and this, together with applicability, will be the most important problems to solve in developing the next version of the system.

With regard to portability, CASE and CHEMICS are written in FORTRAN, and CHEMICS is now running on three different machines.

## REFERENCES

- 1 See, e.g., (a) W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde (Eds.), *Computer Representation and Manipulation of Chemical Information*, Wiley, New York, 1973.
- (b) R. E. Christoffersen (Ed.), *Algorithms for Chemical Computation*, ACS Symposium Series No. 46, 1977.
- (c) D. H. Smith (Ed.), *Computer-Assisted Structure Elucidation*, ACS Symposium Series, No. 54, 1977.
- (d) J. Bargon (Ed.), *Computational Methods in Chemistry*, Plenum, New York, 1980.
- 2 See, e.g., C. H. Davis and J. E. Rush, *Information Retrieval and Documentation in Chemistry*, Greenwood Press, Westport, 1974; Y. Kudo, T. Yamasaki, and S. Sasaki, *J. Chem. Doc.*, 13 (1973) 225.
- 3 A. T. Balaban (Ed.), *Chemical Application of Graph Theory*, Academic, London, 1974.
- 4 K. Nagano, *How to Simulate Protein Folding with Interactive Computer Graphics*, International Summer School on Crystallographic Computing, Prague, 1975.
- 5 Y. Yoneda, CHEMOGRAM, Maruzen, Tokyo, 1972.
- 6 E. A. Feigenbaum and J. Feldman (Eds.), *Computers and Thought*, McGraw-Hill, New York, 1963.
- 7 H. E. Dayringer and F. W. McLafferty, *Org. Mass Spectrom.*, 11 (1976) 895.
- 8 K. Tanabe and S. Sasaki, *Anal. Chem.*, 47 (1975) 118.
- 9 B. A. Jezl and D. L. Dalrymple, *Anal. Chem.*, 47 (1975) 203.
- 10 R. Schwarzenbach, J. Meili, H. Koenitzer, and J. T. Clerc, *Org. Magn. Reson.*, 8 (1976) 11
- 11 S. Sasaki, H. Abe, K. Saito, and Y. Ishida, *Bull. Chem. Soc. Jpn.*, 51 (1978) 3218.
- 12 G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Am. Chem. Soc.*, 90 (1969) 7740.
- 13 E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations*, McGraw-Hill, New York, 1975.
- 14 H. Kuroda and T. Kunii, *Theor. Chim. Acta*, 7 (1967) 220.
- 15 L. P. Lindeman and J. Q. Adams, *Anal. Chem.*, 43 (1971) 1245.
- 16 G. Beech, R. T. Jones, and K. Miller, *Anal. Chem.* 46 (1974) 714.
- 17 L. A. Gribov, M. E. Elyashberg, and V. V. Serov, *Anal. Chim. Acta*, 95 (1977) 75.
- 18 H. B. Woodruff and M. E. Munk, *J. Org. Chem.*, 42 (1977) 1961.
- 19 N. A. B. Gray, *Anal. Chem.*, 47 (1975) 2426.
- 20 B. Pettersson and R. Ryhage, *Anal. Chem.*, 39 (1967) 790.
- 21 A. Buchs, A. B. Delfino, A. M. Duffield, C. Djerassi, R. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, *Helv. Chim. Acta*, 53 (1970) 1394.
- 22 H. Abe and P. C. Jurs, *Anal. Chem.*, 47 (1975) 1829.
- 23 D. B. Nelson, M. E. Munk, K. B. Gash, and D. L. Herald, Jr., *J. Org. Chem.*, 34 (1969) 3800.
- 24 Y. Kudo and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 16 (1976) 43.
- 25 L. M. Masinter, N. S. Sridharan, R. E. Carhart, and D. H. Smith, *J. Am. Chem. Soc.*, 96 (1974) 7714.
- 26 C. A. Shelley, H. B. Woodruff, C. R. Snelling, and M. E. Munk, in D. H. Smith (Ed.), *Computer-assisted Structure Elucidation*, ACS Symposium Series, No. 54 (1977), p. 92.
- 27 R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, 92 (1970) 5755.
- 28 S. Sasaki, I. Fujiwara, H. Abe, and T. Yamasaki, *Anal. Chim. Acta*, 122 (1980) 87.

## CASE, A COMPUTER MODEL OF THE STRUCTURE ELUCIDATION PROCESS

CRAIG A. SHELLEY\*

*Research Laboratories, Eastman Kodak Company, Rochester, NY 14650 (U.S.A.)*

MORTON E. MUNK

*Department of Chemistry, Arizona State University, Tempe, AZ 85281 (U.S.A.)*

(Received 23rd January 1981)

### SUMMARY

CASE (computer-assisted structure elucidation) is an outgrowth of an effort to develop a computer model of structure elucidation. The current capability of CASE in assembling molecules compatible with the structural information entered and in providing interactive user access to the file of assembled molecules is discussed. Of central importance to the effective application of the system to real-world structure problems are the design of the communication link that accepts structural-constraints information and the efficiency of molecule assembly. The chemist—computer interface is designed to accommodate the great and ever-increasing diversity of available structural information. Molecule assembly proceeds by stepwise bond making. Efficient program execution requires the evaluation of each node in the tree against the structural constraints imposed. However, program response time can be improved by heuristic implementation of constraints during molecule assembly and prospective pruning of unproductive branches from the search space as early as possible. CASE must perceive the significance, not only of the individual constraints themselves, but of their interaction as well. A newly added program to edit, analyze, and peruse the constructed list of molecules provides invaluable guidance in the design of the new experiments required to limit further the number of valid candidates. Such fine tuning of the CASE program enhances its utility and ensures that no plausible structures escape consideration.

One of the important methods of determining the structure of an unknown organic compound depends on analysis of its spectral and chemical properties. Three important components of that method are spectrum interpretation, molecule assembly, and spectrum simulation. These same three components form the framework of the computer model of this method, called CASE, an acronym for computer-assisted structure elucidation.

The flow diagram in Fig. 1 provides an overview of the CASE system. To the three major program modules (INTERPRET, ASSEMBLE, and SIMULATE), an EDITOR has been added to facilitate the examination of the output of CASE. The chemist plays a prominent role, as CASE is designed primarily to accelerate structure elucidation and make its conclusions more reliable. This paper describes the user interfaces to ASSEMBLE and EDITOR and some details of the implementation of ASSEMBLE.

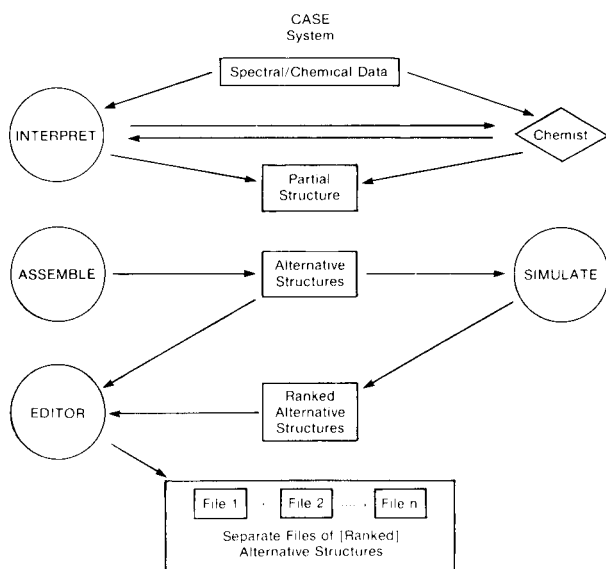


Fig. 1. CASE system flow diagram.

## INTERFACES

### *Interface to ASSEMBLE*

ASSEMBLE is at the heart of the CASE system. It accepts as input computer- and chemist-generated structural information and constructs all structural isomers compatible with this information. The output of ASSEMBLE is a non-redundant listing of compatible molecules in conventional structural language.

Structural information available at any given stage in the structure proof is normally expressed as a partial structure. Therefore, it seemed logical that the input to ASSEMBLE should be a partial structure. For present purposes, the partial structure consists of three components: (1) the molecular formula (elemental composition); (2) non-overlapping structural fragments known to be present in the unknown; and (3) constraints, i.e., the supplementary information that cannot be expressed as non-overlapping fragments; for example, the presence of two cyclopropyl hydrogens in the unknown.

ASSEMBLE elaborates structural isomers, not stereoisomers, i.e., it is concerned only with the topological properties of molecules. Very simply, its role is to expand the partial structure into all complete molecules compatible with it. In the process, bonds are formed between valence-deficient sites in the fragments. The fragments are those entered by the user and all unaccounted-for atoms, which are represented internally as valence-deficient, one-atom fragments.

The constraints on molecule assembly take two forms. They can be local, illuminating the local environment of an atom in a fragment, or they can be global, characterizing the molecule as a whole.

The molecular formula is entered in the standard format (Fig. 2). The molecular formula is obligatory. For entry of non-overlapping structural fragments, a linear code was devised that closely resembles conventional structural language. Thus, a 1-hydroxy-2-methylpropyl group is entered as it would be written (Fig. 2), the residual valence at the 1- and 2-positions being read by the program. Cycles can be conveniently expressed with the same linear code by labeling atoms and then referring to them thereafter by their labels. A benzene ring is easily designated in this fashion as shown by the amino-benzene moiety in Fig. 2.

As indicated, constraints can be either local or global. The local constraints take the form of atom tags. Atom tags, called by the brackets  $\langle$  and  $\rangle$ , provide additional information about the local environment of the atom preceding the tag. Recall that in structural fragment input, an atom of one fragment must not duplicate that of another fragment. Also, the information contained in an atom tag is not counted as part of the fragment. Before this point is elaborated, the available local constraints will be described.

The neighboring-atom tag describes an immediately contiguous atom in terms of the bond by which it must be joined, the element type and hydrogen multiplicity, the hybridization of the atom (i.e., its coordination number),

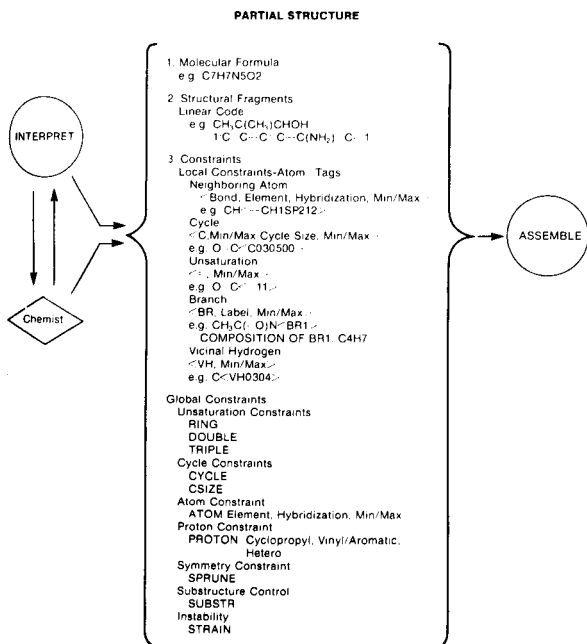


Fig. 2. Interface to ASSEMBLE.

and finally, the minimum and maximum number of such contiguous atoms to be permitted. In the example shown in Fig. 2, the fragment is a carbon atom bearing one hydrogen. The information in the tag requires the contiguous atom to be carbon, to join to the tagged atom by a single bond, to bear a single hydrogen, and to be tricoordinate. The minimum and maximum values require at least one such contiguous atom, but no more than two.

The cycle tag, called by the symbol C, sets restrictions on the presence or absence of the tagged atom in a cycle or cycles of specified size. For example, if the presence of an unstrained carbonyl is inferred from the i.r. spectrum, the cycle tag gives that information. Note that the tagged atom (Fig. 2) is the carbonyl carbon. The tag forbids the presence of the carbonyl carbon atom in a 3- to 5-membered cycle; C to call the cycle tag, 03 for minimum cycle size, 05 for maximum cycle size, 0 for the minimum number of such cycles permitted, 0 for the maximum number permitted.

The unsaturation tag, called by the symbol =, designates the presence or absence of  $\alpha,\beta$ -unsaturation to the tagged atom. In the simple example (Fig. 2), valid molecules require a carbonyl group with a single occurrence of  $\alpha,\beta$ -unsaturation.

With the branch tag, the presence of a group at any atom in a fragment may be specified in terms of the atom composition of the group. This kind of information is often derived from high-resolution m.s. data. The tag is called by the symbol BR, followed by the branch label (Fig. 2). As shown in Fig. 2, the user is then asked to enter the atom composition of the branch. Here the branch joins to amide nitrogen. Because the optional minimum and maximum values are not included, at least one branch of composition  $C_4H_7$  must join to nitrogen.

The *vicinal* hydrogen tag, denoted by the symbol VH, can express the presence or absence of hydrogen atoms on atoms contiguous to the tagged atom. An exact number or a permissible range may be specified by means of the minimum and maximum values. In the example (Fig. 2), at least three, but no more than four, hydrogen atoms are to be permitted on atoms contiguous to the tagged carbon atom.

In addition to illuminating the local environment of atoms, these local constraints are a useful device in resolving, in part, the ambiguities associated with fragments that may overlap one another. Since the information contained in atom tags is not counted as part of the fragment, that information cannot lead to overlapping fragments. For example, if it is possible that an atom in one fragment may duplicate an atom in another fragment, that atom may be expressed by means of the neighboring-atom tag in one of the two fragments.

Consider the example shown in Fig. 3. A naturally occurring toxin was shown to have a cyclopentanone unit in which the ketone carbonyl was unconjugated and adjacent to a methylene group and a methine group. The fragment inferred is shown on the left. Note that the input statement

FRAGMENT

INPUT

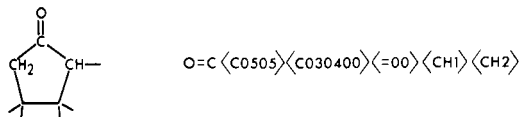


Fig. 3. The use of atom tags to avoid the overlap problem.

on the right need not include all atoms of the 5-membered ring. The cycle tag, <C0505>, requires the ketone carbonyl to be part of a 5-membered ring. Thus, the four remaining non-hydrogen atoms of the ring, which conceivably may duplicate atoms in other fragments, need not be included in the input statement. The methylene and methine carbon atoms adjacent to the ketone carbonyl can be expressed as neighboring-atom tags, <CH1> and <CH2>. The cycle tag, <C030400>, excludes the ketone carbonyl from a 3- or 4-membered ring. The unsaturation tag, <=00>, precludes an  $\alpha,\beta$ -unsaturated ketone. Thus, the fragment shown in Fig. 3 can be represented by using a simple carbonyl-group structural fragment and augmenting it with atom tags if there is a possibility that atoms in this fragment may duplicate atoms in other fragments.

The global constraints are designed with a heavy emphasis on the application of spectral data (Fig. 2). First, with the unsaturation constraints, RING, DOUBLE, and TRIPLE, the unsaturation equivalents can be specified in terms of the number of rings and/or the number of double and triple bonds. The "number" specified may be an exact number or a permissible range.

Second, the constraints controlling cycles are CYCLE and CSIZE. The former expresses the number of cycles permitted, the latter the number of cycles of specific cycle size. Again, the number of cycles and the size of cycles can be expressed as an exact number or a permissible range.

Third, the number of atoms of a particular coordination number and/or a particular hydrogen multiplicity is often known from n.m.r. studies. The ATOM constraint inputs such information.

Fourth, n.m.r. may likewise identify the presence of particular types of hydrogen atoms, e.g., cyclopropyl hydrogens, aromatic and vinyl hydrogens, and hydrogens on carbon bearing electronegative heteroatoms. PROTON allows the input of this information.

Fifth, if the SPRUNE constraint is called, the number of signals expected in the <sup>13</sup>C-n.m.r. spectrum of each molecule constructed by ASSEMBLE is predicted. A structure is either retained on the file of valid molecules or deleted in accord with user instructions on the closeness of the fit required between the predicted and the observed number of signals. Since the number of signals in the <sup>13</sup>C-n.m.r. spectrum of a given compound is a measure of its symmetry, SPRUNE in effect constrains the list of valid molecules on the basis of symmetry properties.



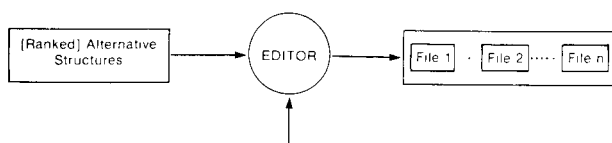
Substructure control is a versatile and powerful constraint. With it the user can require the presence or the absence of any defined substructure. An example of the versatility of substructure control is the "periodate substructure." A set of rules defining the cleavage of the carbon-carbon single bond by periodic acid can be entered by using substructure control. Now the user need only enter the number of molecules of periodate consumed by the unknown. Only those molecules compatible with this value will be assembled.

Finally, the routine STRAIN operates automatically without being called and perceives the formation of several commonly encountered sources of molecular strain that leads to high instability. Such molecules never appear in the file of valid molecules unless the STRAIN routine is disengaged by the user.

The output of ASSEMBLE, i.e., a file of valid molecules, may be systematically organized and, if desired, pruned, according to a broad range of instructions written into the EDITOR program. A systematic presentation of the valid molecules facilitates the design of the most direct experimental strategy to narrow the list of candidates to one. In fact, at times an examination of the range of structure types constructed by ASSEMBLE promotes the intuitional leaps that dramatically shorten the process leading to the correct assignment.

### Interface to EDITOR

The global constraints in ASSEMBLE also form the basic core of the instructions that control program EDITOR (Fig. 4). Thus, separate files can



#### FILE CREATION COMMANDS

##### 1. Molecular Skeleton

###### SKELETON

HETERO  
BRANCHES  
ACYCLIC

###### SKELETON FILE

##### 2. Constraints

RING  
DOUBLE  
TRIPLE  
CYCLE  
CSIZE  
SUBSTR  
ATOM  
PROTON  
CYCLOPROPYL  
VINYL AROMATIC  
HETERO  
STRAIN

##### 3. Spectral Properties

CARBON-13

##### 4. Logical Operations

AND  
OR  
NOT

##### 5. General Commands

DELETE  
NUMBER  
DRAW

Fig. 4. Interface to the EDITOR.

be created on the basis of a user-defined number of rings or multiple bonds (RING, DOUBLE, and TRIPLE), the number of cycles (CYCLE), the number of cycles of specified size (CSIZE), the number of atoms of specific element type, hydrogen multiplicity, and coordination number (ATOM), the number of protons of specified type, e.g., cyclopropyl, vinyl/aromatic, or those attached to carbon bearing heteroatoms (PROTON), and the number of user-defined substructural fragments (SUBSTR). If the strain routine was disabled during molecule assembly, then a separate file of strained molecules can be created (STRAIN).

The capability to organize files on the basis of skeletal type is included in the EDITOR. Three different options are provided. HETERO generates a non-redundant list of skeletons by removing all non-carbon atoms and then designating the largest fragment remaining as the skeleton; BRANCHES generates skeletons by removing terminal atoms from molecules until no terminal atom remains; ACYCLIC produces skeletons by pruning all atoms that are not part of a cycle. In all three routines, only connectivity is preserved in the skeleton, i.e., multiple bonds are replaced by single bonds. The SKELETON FILE command creates a file of complete molecules having a particular skeletal type in common.

CARBON-13 creates files of compounds on the basis of the predicted number of  $^{13}\text{C}$ -n.m.r. signals. Finally, any combination of subfiles can be created by the Boolean commands, AND, OR, and NOT.

Subfiles can be deleted (DELETE). The number of compounds in any given subfile can be requested (NUMBER), and the compounds in one or more subfiles can be presented to the user in conventional structural language (DRAW).

#### IMPLEMENTATION OF ASSEMBLE

In its execution, ASSEMBLE uses a depth-first search to expand the partial structure to the complete list of compatible molecules. A trace of the program execution is illustrated by the "tree" shown in Fig. 5. Each non-terminal node in the search tree is an incomplete molecule. The "root" node in the tree, represented by a square, is the partial structure entered by the user. Each terminal node, represented by a triangle, is a complete molecule. Each edge represents the pathway traced by making one or more "connec-

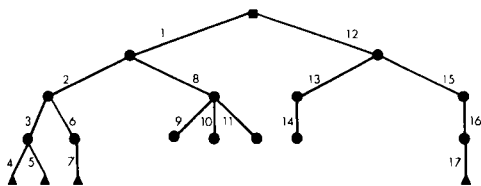


Fig. 5. Depth-first search.

tions" between two atoms. Note that multiple bonds in ASSEMBLE are represented by multiple connections between two atoms in the connection table. The integer labels on the edges in Fig. 5 illustrate the depth-first order in which the tree is traversed during molecule assembly. It should be apparent that the major component of the construction procedure is node selection. "Intelligent" node selection is vital; without it, ASSEMBLE would continually be entering branches in the tree (Edge 8, Fig. 5) that would not lead to complete molecules or only duplicate work performed in a previous branch. In real-world problems such a deficiency could lead to the use of prohibitive amounts of computer time.

Node selection consists of two important procedures: first, evaluating each node for compatibility with the imposed constraints and, second, perceiving symmetry relationships to avoid duplicating previous steps. The second procedure, although important, will not be covered in this paper.

ASSEMBLE is intended to be conversational. Therefore, every attempt is made to use the constraint information contained in the partial structure such that program efficiency is maximized. Except for SPRUNE, all fragments, local constraints, and global constraints are currently used to constrain molecule assembly.

Now consider how the cycle tag might be implemented to provide insight into constraint implementation in ASSEMBLE. Recall that the cycle tag consists of a mnemonic, C, the minimum and maximum cycle size, and an optional minimum and maximum on the number of cycles in the range that contain the tagged atom. The tag <C0505> would require the tagged atom to be in at least one 5-member cycle.

The following problem is illustrative. The only information entered consists of the molecular formula,  $C_6H_{10}O$ , and one fragment, a carbonyl with a cycle tag indicating that at least one 5-member cycle contains the carbonyl carbon ( $O=C\langle C0505 \rangle$ ). Since this cycle tag does not have an effective maximum, only the minimum needs to be checked. During molecule assembly, the cycle tag could be checked for consistency with each prospective connection. Consider the following heuristic for the implementation of the constraint minimum: if all potential ring unsaturations have not been used by making a connection or the tagged atom is in a cycle of the designated size, then accept the proposed connection; otherwise, reject the proposed connection.

With this heuristic, the molecule assembly tree shown in Fig. 6 will be traced. Note that the molecular formula requires two unsaturation equivalents. Within ASSEMBLE, the atoms are assigned sequence numbers at the root node in the tree as shown in Fig. 6; the carbonyl carbon is assigned sequence number 1 and the other carbons are assigned sequence numbers 2-6. The first connection is from atom 1 to atom 2. Note that potential connections from atom 1 to atoms 3, 4, 5, or 6 at the root node are recognized as symmetry-forbidden by ASSEMBLE. The next connection is between atoms 1 and 3 to form the ketone moiety. Now a connection is attempted

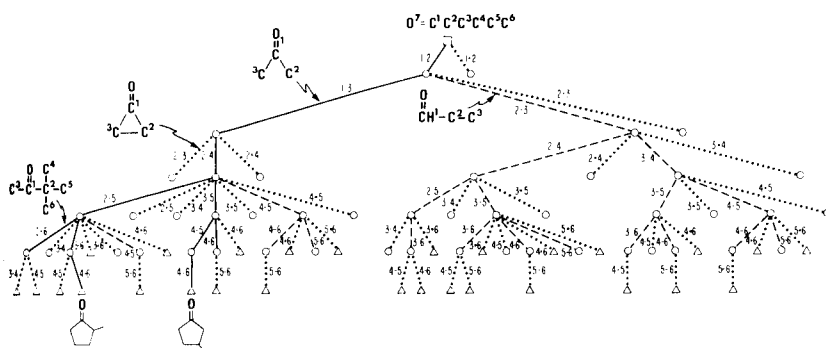


Fig. 6. A molecule assembly tree.

from atom 2 to atom 3, which forms a 3-member cycle. Since all potential ring unsaturations have been used and the tagged atom is not contained in a 5-member cycle, this proposed connection is rejected according to the heuristic. Next, ASSEMBLE backs up to the parent node in the tree and makes the next possible connection, atom 2 connected to atom 4. Subsequent connections are formed between atoms 2 and 5 and then 2 and 6. The connection from atom 2 to atom 6, however, cannot lead to complete molecules, because it forms a fragment that is too highly branched for a cycle of the required size to be formed. ASSEMBLE, however, would not recognize this, using the previously stated heuristic, and would continue its depth-first search down the tree. Both of the connections 3 to 4 and 4 to 5 lead to cyclic molecules that do not satisfy the cycle tag heuristic and are consequently rejected. Now ASSEMBLE backtracks to the parent node of the 2 to 6 connection. Since atom 2 cannot attach to any other non-hydrogen atom, ASSEMBLE bonds all the free valences on atom 2 to hydrogen atoms. A connection is now proposed between atoms 3 and 4 but it is rejected with the cycle tag heuristic. The next connection, 3 to 6, eventually leads to a complete molecule, 1-methylcyclopentanone, that satisfies the input constraints.

Dotted edges in the tree represent proposed connections that are rejected by the cycle tag heuristic. All other edges were accepted by ASSEMBLE using this heuristic. It is apparent that a rather large branch of the tree is entered with the 2 to 3 connection following the initial 1 to 2 connection that does not produce any molecules consistent with the constraint. At the parent node of this connection, all possible connections between atom 1 and other non-hydrogen atoms have been previously attempted. Consequently, the program bonds all the free valences of atom 1 to hydrogen atoms before making the 2 to 3 connection to produce the aldehyde fragment shown. Clearly, no molecules consistent with the cycle tag constraint can be produced from this fragment. Note that ASSEMBLE does not explicitly represent hydrogen atoms in the connection table. Therefore, rather than attaching hydrogen atoms to atom 1 in the connection table, atom 1 is

simply flagged to designate that no additional free valences on this atom can bond to non-hydrogen atoms below this node in the tree.

As ASSEMBLE has evolved, inefficient heuristics such as this cycle tag heuristic have been replaced by substantially more effective ones. The current heuristic for cycle tag minimum is stated as follows: (1) count the number of cycles that currently satisfy the constraint; (2) if the tag minimum is already satisfied, then accept the proposed connection; (3) if all ring saturations have been used, then reject the proposed connection; (4) if the tagged atom has at least two free valences, then accept the proposed connection; (5) if the tagged atom is contained in a path that is terminated by atoms with free valence and the length of this path is not longer than the maximum cycle size, then accept the proposed connection. Otherwise, reject the proposed connection.

By applying this heuristic to the tree, all dashed edges are removed. Thus, only the edges represented as solid lines are accepted during constraint evaluation. In other words, this improved heuristic reduced the number of edges accepted by ASSEMBLE from 30 to 11, substantially improving program efficiency for this problem. In practice, not only do improved heuristics produce substantial efficiency gains, but they also transform impractical problems, with unrealistic c.p.u. time requirements, to problems that can be solved with reasonable c.p.u. demands.

Although individual constraints often require substantial effort to implement them efficiently, far more problems are encountered with real-world problems, where the interplay between constraints can be crucial to program efficiency. Consider the following problem. The molecular formula is  $C_{12}H_{22}O_2$ . One fragment is the alcohol moiety (OHC<C33>). The atom tag <C33> requires three carbon atoms to be attached to the tagged atom; in other words, the alcohol must be tertiary. The second fragment (O=CH<CH1SP2>) is an aldehyde group that is required by the atom tag to be connected to an  $sp^2$ -hybridized methine carbon. Now assume that ASSEMBLE attempts to connect these fragments. For the connection OHC<C33>CH<CH1SP2>=O, the first carbon atom is required to be attached to two additional carbons and to be  $sp^3$ -hybridized by the tag <C33>, and at the same time, the tag <CH1SP2> requires the same atom to be  $sp^2$ -hybridized. If all tags were implemented independently by ASSEMBLE, this inconsistency would not be detected by ASSEMBLE at this time and the proposed connection would be accepted. This connection would lead to extreme inefficiency. However, ASSEMBLE does detect the problem by flagging atoms with free valence where neighboring-atom tags restrict multiplicity and/or hybridization on the same atom.

## A COMPUTER SEARCH SYSTEM FOR CHEMICAL STRUCTURE ELUCIDATION BASED ON LOW-RESOLUTION MASS SPECTRA

K. S. LEBEDEV, V. M. TORMYSHEV, B. G. DERENDYAEV and V. A. KOPTYUG\*

*Scientific Information Centre on Molecular Spectroscopy, Novosibirsk Institute of Organic Chemistry, Siberian Division of the USSR Academy of Sciences, Prospekt Nauki 9, 630090 Novosibirsk (U.S.S.R.)*

(Received 23rd January 1981)

### SUMMARY

A computerized search system which employs the data on the masses and relative abundances of spectral peaks and primary neutral losses is designed for computer elucidation of chemical structures. Recognition of structural fragments is based on analysis of the structures of reference compounds selected as best matches to the mass spectrum of the compound under investigation. Tests of the system on 67 "unknowns" show that the probability of recognizing a large structural fragment lies in the interval 60–80%, depending on the fragment size (100–50% of molecular weight), and that the reliability of the corresponding structural conclusion is 98%. An approach to automatic selection of the substructure common to all or several of the selected compounds is discussed.

Mass spectrometry and other methods of molecular spectroscopy are widely used in the elucidation of organic compound structure. However, interpretation of mass spectra faces serious difficulties arising from the complicated relationship between molecular structure and the nature of molecular fragmentation under electron impact. Computerized systems developed to solve this problem can be of substantial help to investigators, and it is not surprising that a great number of papers on this subject has appeared during the last fifteen years. The systems reported for low-resolution mass spectrometry can generally be classified as retrieval systems and spectral interpretation systems.

Retrieval systems [1–6] are designed to identify previously described compounds by consecutively matching the mass spectrum of the compound examined against reference spectra stored in the computer file. These systems are employed in the examination of environmental pollution, in forensic work and in some other fields.

Spectral interpretation systems are designed to assist investigators in structure elucidation of unknown compounds. Three main trends can be distinguished in these studies.

*Pattern recognition* [7–9]. This method enables conclusions to be drawn on the presence or absence of certain structural features in the molecule of a compound under investigation; consequently, the unknown compound can

be attributed to a definite chemical class. The classification is based on the decision function formulated for every specific substructure on the basis of the preliminary training session, which requires a large investment of computer time and human effort. This limits the applicability of the method for solving structural problems of arbitrary character. However, it is appropriate to use such systems in the final stage of molecular structure elucidation, when it is necessary to choose between alternative hypotheses.

*Artificial intelligence* [10–13]. This type of system is characterized by preliminary input into the computer memory of empirically established rules of molecular fragmentation, and by the use of these rules for the analysis of spectra of unknown compounds, just as an investigator would do it. This method appears to suffer both from the complicated relationships that exist between structure and mass spectra, and from the considerable effort necessary to develop subprograms for each individual chemical class of compounds. The artificial intelligence approach requires that fragmentation rules be known before spectral interpretation. This is a serious limitation and therefore the design of a universal interpretive system based on artificial intelligence appears to be difficult, if possible at all. Such systems have been developed for some classes of compounds, e.g., aliphatic ketones [11], amines [12] and alcohols [13].

*Search for structural analogs* [14, 15]. Systems of this type involve selection from the reference file of a number of spectra which are closely similar to the test spectrum, and so indicate some compounds with close structural relationship to the unknown. This approach does not place any restrictions on the classes of compound examined, provided that spectra of the various types of organic compounds are stored in the computer file.

To develop an efficient computerized system for mass spectrometry, it seems advisable to combine the advantageous features of retrieval systems and structural analogs, and, if necessary, to utilize pattern recognition and artificial intelligence. Such a system is now being developed at this Centre.

## EXPERIMENTAL

To compile the reference spectra library, it is advisable to condense the original mass spectra in some standardized manner in order to reduce both the computer storage requirement and the operational time of the system. One of the widely used methods for abbreviation of mass spectra is to select one or two of the most intense peaks from each interval of 14 mass units throughout the spectrum [1]. However, a statistical study of our reference file of 23000 mass spectra shows that the probability of peaks occurring varies for different mass regions. The average probability values for the regions of 20–117  $m/z$ , 118–299  $m/z$  and 300–705  $m/z$  are close to  $(1/2)^{2.5}$ ,  $(1/2)^4$  and  $(1/2)^8$ , respectively. This means that the peaks to be retained for each 14 mass unit interval in the regions mentioned should be in the ratio 3:2:1 [16, 17]. An example of the original and abbreviated spectra is shown in Fig. 1.

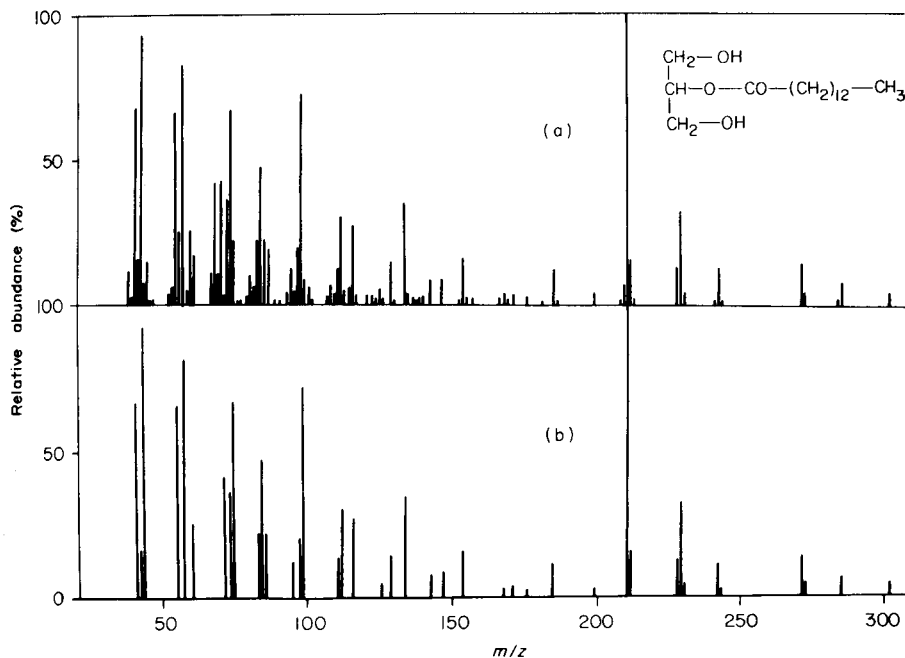


Fig. 1. Examples of the original (a) and compressed (b) mass spectra of 2-monomyristin.

Further, a statistical analysis of peak intensity distribution defines the five possible windows of abundance values: 1–8.3%, 8.4–24.2%, 24.3–50.2%, 50.3–87.7%, and 87.8–100%. The probabilities ( $P_i$ ) of a randomly chosen peak intensity appearing in these windows are  $1/2$ ,  $(1/2)^2$ ,  $(1/2)^3$ ,  $(1/2)^4$  and  $(1/2)^5$ , respectively. Thus, five intensity gradations  $I_i = -\log_2 P_i$  (with  $I_i = 1, 2, 3, 4, 5$ ) were used for intensity representation in the abbreviated reference mass spectra.

The probability of peak occurrence at a certain mass position decreases with increasing mass number. Accordingly, when the similarity of reference spectra to the spectrum of the unknown compound is evaluated, higher significance is attributed to “unique” high peak masses in accordance with the formula  $M_i = -\log_2 P_i(m/z)$ , where  $P_i(m/z)$  is the probability of peak occurrence at a specified mass position  $m/z$ . A similar approach is taken with peak intensity, i.e., the higher the intensity, the greater the significance.

To find the best matches arising from comparison of the unknown spectrum (x) with those from the computer catalog (c) when the system is run in the identification option, the unit match factor  $W_i^x$  is calculated for each peak of the unknown spectrum from the equations

$$W_i^x = M_i^x + I_i^x, \text{ if } I_i^x = I_i^c \quad (1)$$

$$W_i^x = M_i^x, \text{ if } I_i^x \neq I_i^c \quad (2)$$

It is implicit that mass positions in the two spectra under comparison co-



incide, otherwise  $W_i^x = 0$ . The overall match factor for the whole spectrum is defined by  $W_A^x = \sum_i W_i^x$ .

The coincidence criterion  $F$  for each pair of spectra is then calculated from

$$F = 100 W_A^x / \sum_i (M_i^x + I_i^x), \quad (3)$$

where the peak mass and peak intensity increments ( $M_i^x$  and  $I_i^x$ ) are summed for all peaks of the unknown spectrum.

The compounds selected as the best matches are arranged in decreasing order of the  $F$  criterion in the computer output. Examples of the machine output are shown in Table 1.

The system efficiency was tested by using a control set of 217 arbitrary spectra. These spectra were collected from the literature and were known to be measurements different from those in the library. It was shown that the probability of occurrence of a correctly identified compound among the top five compounds of the search output is approximately 96% and increases to 99% if the molecular weight is specified.

Another option designed for retrieval of compounds closely similar in structure to the unknown uses a somewhat modified library of reference spectra. Nevertheless, its distinctive features are not so significant as to require special comment. This option tries to use not only the absolute peak positions in the pair of reference and unknown spectra being compared, but also their relative positions in order to reveal characteristic ion series corresponding to primary losses with the same masses (Fig. 2). The significance of the match of the primary losses ( $M - m/z$ ) occurring in both unknown and reference spectra is evaluated by increments  $L_i = -\log_2 P_i(M - m/z)$ ,

TABLE 1

Examples of the computer outputs obtained for mass spectra of a number of unknowns (only the three top compounds are presented)

Unknown compound	$F(\%)$	Top three compounds of computer answer
HEXACHLOROETHANE	75	HEXACHLOROETHANE
	37	1,1-DIFLUOROPERCHLOROPROPANE
	34	METHYL TRICHLOROACETATE
5-PHENYLEICOSANE	68	5-PHENYLEICOSANE
	61	11-PHENYLEICOSANE
	61	7-PHENYLEICOSANE
2-HEPTANOL	73	2-HEPTANOL
	65	3-METHYL-2-HEXANOL
	58	2-OCTANOL
O-METHYLBENZYL ALCOHOL	77	O-METHYLBENZYL ALCOHOL
	39	BETA-PHENYLETHYL FORMATE
	36	1-PHENYLETHYL ACETATE

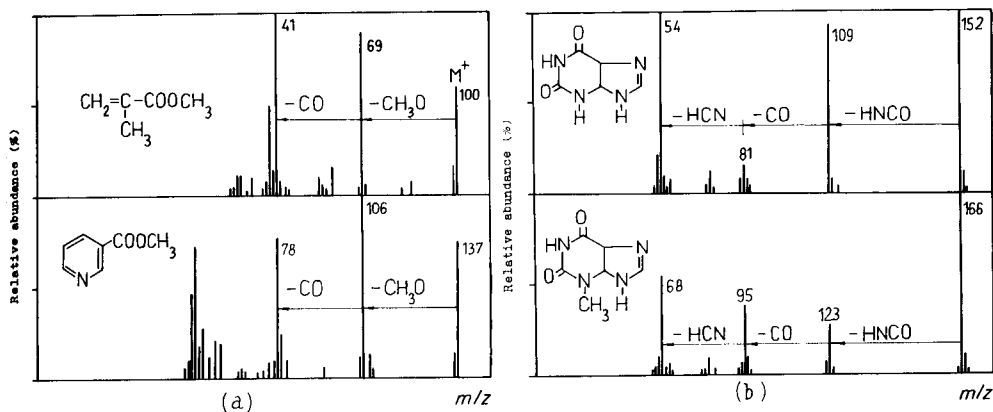


Fig. 2. Possible types of fragmentation of structurally related molecules: (a) common fragment losses from the substituent; (b) common fragment losses from molecule skeleton.

where the probabilities of neutral primary losses  $P_i(M - m/z)$  result from statistical analysis of the mass spectra library. The overall match factor  $W_B^x$  which takes into account the coincidence of both the relative peak positions and the corresponding peak intensities is calculated similarly to the factor  $W_A^x$  (see above), with parameters  $M_i$  replaced by  $L_i$ .

Finally, the overall match factor  $W_{AB}^x$  is calculated from the equation  $W_{AB}^x = W_A^x + W_B^x$  for two categories of data: absolute peak positions and primary neutral losses.

## RESULTS AND DISCUSSION

Table 2 demonstrates the top four compounds presented by the system as the computer output for a number of search requests which were implemented

TABLE 2

Examples of the top four compounds selected by the use of match factors  $W_{AB}$

Unknown compound	$W_{AB}$	Top four compounds of computer answer
2-N-BUTYLTHIOPHENE	68	3-TERT-BUTYLTHIOPHENE
	62	2-N-PROPYLTHIOPHENE
	61	2-N-HEXYLTHIOPHENE
	56	2-ETHYLTHIOPHENE
2,5-DIMETHYL-3-N-PROPYLPYRAZINE	98	2,3-DIMETHYL-5-N-BUTYLPYRAZINE
	91	2,3-DIMETHYL-5-(2-BUTYL)-PYRAZINE
	89	2,5-DIMETHYL-3-(3-METHYLBUTYL)-PYRAZINE
	87	2,5-DIMETHYL-3-ISOBUTYLPYRAZINE
DI-P-TOLYL ETHER	129	O-TOLYL-P-TOLYL ETHER
	85	PHENYL-M-TOLYL ETHER
	79	PHENYL-O-TOLYL ETHER
	74	AR-ETHYLPHENYLPHENYL ETHER

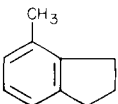
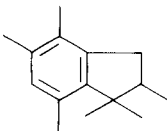
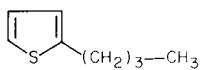
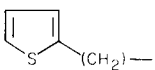
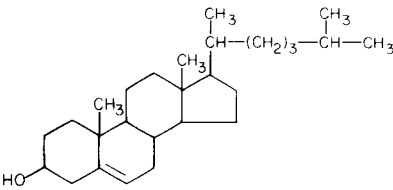
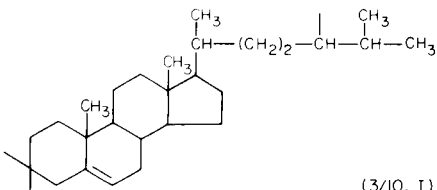
by the combined technique mentioned above (the primary losses were examined in the mass range from 0 to  $M/2$ ). To test the ability of the system to retrieve from the computer library the spectra of compounds with structures closely related to the "unknown", three conditional categories of macrofragments, resulting from comparison of the computer output compounds, were determined. Categories I, II and III correspond, respectively, to 100–75%, 75–60% and 60–50% of the original molecule size. The calculation of the macrofragment (or molecule) size takes into account all types of atoms except hydrogens.

The system was examined for 67 mass spectra of a wide variety of organic compounds. To make the macrofragment prediction, the structures of the top compounds of every search output were compared in order to identify the largest substructure common to several selected compounds [17] (for examples of the macrofragments obtained by this technique, see Table 3).

TABLE 3

Examples of macrofragments recognized by comparison of structures of the computer output

(The values in parentheses indicate the ratio of the number of macrofragment occurrences to the number of computer output compounds used for macrofragment prediction, and the size categories of the macrofragments.)

Test compound	Revealed macrofragment	
$\text{CH}_2=\text{CH}-(\text{CH}_2)_4-\text{CH}_3$	$-\text{CH}=\text{CH}-(\text{CH}_2)_4-$	(14/14, I)
$\text{CH}_3-(\text{CH}_2)_5-\text{COOCH}_3$	$-(\text{CH}_2)_6-\text{COOCH}_3$	(14/14, I)
		(6/6, I)
		(11/14, I)
		(3/10, I)

The probabilities of correct macrofragment prediction were found to be 61, 72 and 84% for categories I, II and III, respectively, when the three "best" compounds from the computer output were used in the procedure mentioned above. The reliability of the corresponding structural conclusions was approximately 98% [18].

It is tempting to turn to the automatic comparison of compounds from the search output in order to facilitate and speed up the procedure of common substructure extraction. To solve this difficult problem, a library of compound structures complementing the mass spectra file was compiled, using the computerized system for coding organic compound structures. Such a system for the structure coding was developed at this Centre [19]. It is noteworthy that this system is based on the atomic-fragment code, which employs widely used microfragments such as  $\text{CH}_3$ ,  $\text{CH}_2$ ,  $\text{CH}$ ,  $\text{OH}$ ,  $\text{NH}_2$ ,  $\text{NO}_2$ , along with the individual atoms. This code provides compact representation of the two-dimensional organic structures in the form of connectivity matrices. However, the original code proved to be insufficiently effective for realization of the computer extraction of common macrofragments from the arbitrary structure set. Thus, for instance, examination of the three compounds in Fig. 3 represented in the original atomic-fragment code (a) gave the maximal common fragment (I), whereas the use of the modified code

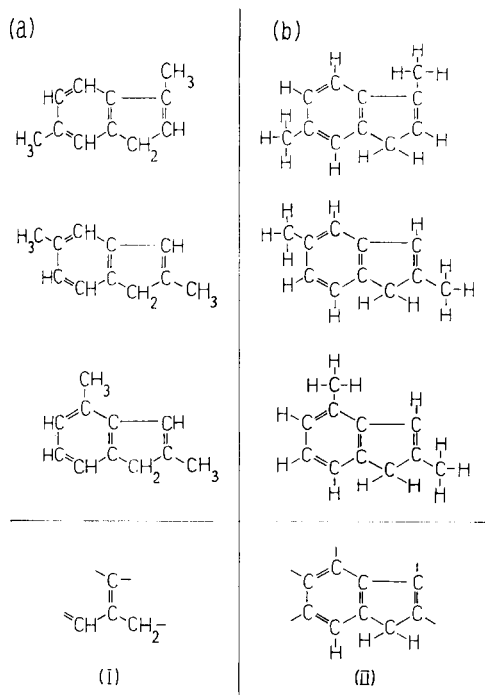


Fig. 3. Fragments common for three structures represented in different codes: (a) atomic-fragment code; (b) atomic code.

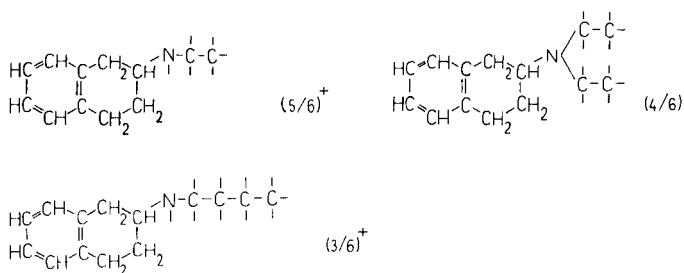


Fig. 4. The fragments revealed from the computer analysis of the top six compounds of the search output for *N*-cyclohexyl-1,2,3,4-tetrahydro-2-naphthylamine as an unknown. (The values in parentheses indicate the number of compounds containing the given fragment in the top six compounds of the computer answer; the "plus" indicates the correct macrofragment.)

providing molecule representation as an atomic graph (b) made it possible to determine the maximal common fragment II with a much larger size. A fragment of similar constitution can also be found from a comparison of structures coded in the skeleton form which can easily be derived from the atomic form by cutting off all the hydrogens. Each method of structure representation has its own indisputable merits and disadvantages which become noticeable in the treatment of concrete structures. For this reason, the approach reported here tries to make manipulation possible with the structural information presented in different forms; original atomic-fragment code, and modified atomic and skeleton codes.

The algorithms and programs developed perform the comparison of the structures selected by the computer as a result of the search procedure, extraction of the fragments which are common to all or some of these structures, and calculation of the number of compounds containing the specified fragments. For example, the computer analysis of the top six compounds selected for the mass spectrum of *N*-cyclohexyl-1,2,3,4-tetrahydro-2-naphthylamine made it possible to produce the structural fragments presented in Fig. 4.

For future development of the system, its conjugation with systems for structure elucidation based on other spectroscopic methods is envisaged, with subsequent amalgamation of all the information originating from a variety of sources in order to generate the most probable structures of the compound being examined.

## REFERENCES

- 1 H. S. Hertz, R. A. Hites and K. Biemann, *Anal. Chem.*, 43 (1971) 681.
- 2 B. A. Knock, I. C. Smith, D. E. Wright and R. G. Ridley, *Anal. Chem.*, 42 (1970) 1516.
- 3 S. L. Grotch, *Anal. Chem.*, 45 (1973) 1.
- 4 S. R. Heller, *Anal. Chem.*, 44 (1972) 1951.
- 5 G. M. Pesyna, R. Venkataraghavan, M. E. Dayringer and F. E. McLafferty, *Anal. Chem.*, 48 (1976) 1362.

- 6 B. G. Derendyaev, L. M. Pokrovsky, S. A. Nekhoroshev, V. I. Smirnov and V. A. Koptyug, *Izv. Sib. Otd. Akad. Nauk SSSR. Ser. Khim. Nauk*, (4) (1977) 109.
- 7 V. L. Talrose and V. V. Raznikov, *Dokl. Akad. Nauk SSSR*, 159 (1964) 182.
- 8 V. V. Raznikov and V. L. Talrose, *Dokl. Akad. Nauk SSSR*, 170 (1966) 379.
- 9 P. C. Jurs and T. L. Isenhour, *Chemical Application of Pattern Recognition*, Interscience-Wiley, New York, 1975.
- 10 B. G. Buchanan, A. M. Duffield and A. V. Robertson, in G. W. A. Milne (Ed.), *Mass Spectrometry: Techniques and Applications*, Interscience-Wiley, New York, 1971, p. 121.
- 11 A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum and J. Lederberg, *J. Am. Chem. Soc.*, 91 (1969) 2977.
- 12 A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum and J. Lederberg, *J. Am. Chem. Soc.*, 92 (1970) 6831.
- 13 B. V. Rozynov, I. A. Bogdanova, G. I. Spivakovskii, I. I. Zaslavskii, A. I. Tishchenko and N. S. Wulfson, *Izv. Akad. Nauk SSSR*, (7) (1974) 1526.
- 14 K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty, *J. Am. Chem. Soc.*, 95 (1973) 4185.
- 15 P. R. Naegeli and J. T. Clerc, *Anal. Chem.*, 45 (1974) 739A.
- 16 B. G. Derendyaev, V. A. Koptyug, K. S. Lebedev and O. N. Sharapova, *Avtometriya*, (4) (1979) 3.
- 17 K. S. Lebedev, V. M. Tormyshev, O. N. Sharapova, N. B. Mamaeva, B. G. Derendyaev and V. A. Koptyug, *Izv. Sib. Otd. Akad. Nauk SSSR. Ser. Khim. Nauk*, (4) (1980) 54.
- 18 K. S. Lebedev, V. M. Tormyshev and B. G. Derendyaev, *Izv. Sib. Otd. Akad. Nauk SSSR. Ser. Khim. Nauk*, (4) (1980) 64.
- 19 V. N. Piottukh-Peletsky, M. I. Podgornaya, V. I. Smirnov, G. G. Balakina and V. A. Koptyug, *Izv. Sib. Otd. Akad. Nauk SSSR. Ser. Khim. Nauk*, (12) (1976) 134.

## COMPUTER-AIDED STRUCTURE ELUCIDATION OF ORGANIC COMPOUNDS WITH THE CHEMICS SYSTEM

### Removal of Redundant Candidates by $^{13}\text{C}$ -n.m.r. Prediction

I. FUJIWARA, T. OKUYAMA, T. YAMASAKI\*\*, H. ABE and S. SASAKI\*

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

(Received 23rd January 1981)

#### SUMMARY

A  $^{13}\text{C}$ -n.m.r. prediction module capable of removing inappropriate candidate structures given for an unknown compound based on the spectral data is introduced for the CHEMICS system. Given a set of candidate structures generated in the system, the routine may be used to prune off redundant candidates which have a predicted number of signals inconsistent with the observed number. It is shown that the addition of the examination module to the system makes structure elucidation by computer much more practical.

The CHEMICS computer program system has been developed to elucidate the structures of organic compounds from spectroscopic data and other structural information [1]. The ultimate goal of this system is to obtain a single correct structure for an unknown compound. In practice, however, a large number of candidate structures is sometimes generated when the system is applied to molecules with large and/or complicated structures. In such cases, simulation of spectral data or prediction of chemical or physical properties is very useful in reducing the number of candidate structures as well as in indicating structural constraints, e.g., insertion of substructures designated by the user at appropriate stages of the computation [2].

In this paper, a function of the module for examination of candidates generated in the system is described. The function is the prediction of the number of signals in a broad-band decoupled  $^{13}\text{C}$ -n.m.r. spectrum from the topological representation of the candidate structure by means of symmetry perception. The predicted number of signals for each candidate compared with the observed number. After structure construction, this option may be used to prune the list of candidates in an interactive mode. This module provides the system with a powerful means of reducing the number of candidate structures.

Shelley and Munk [3] have reported a program for prediction of the number of signals in a  $^{13}\text{C}$ -n.m.r. spectrum. The basic idea of the present

\*\*Present address: Mitsui Petrochemical Industry, Co. Ltd., Iwakuni, Yamaguchi 740, Japan.

work is rather similar to their approach, which is based on the following assumption. In general, the number of signals in a broad-band decoupled  $^{13}\text{C}$ -n.m.r. spectrum may be equal to the number of structurally nonequivalent carbon atoms. Therefore, as a first approximation, it can be postulated that the number of topologically nonequivalent carbon atoms is equal to the number of signals in a  $^{13}\text{C}$ -n.m.r. spectrum [4].

It is well known that topologically identical carbon atoms are not always structurally equivalent because of the stereochemical condition of the organic compound. Thus, as pointed out by Shelley and Munk, some modifications must be considered for precise prediction of the signal number (See EXPERIMENTAL).

The perception of topologically distinct carbon atoms is just the same as that of equivalent carbon atoms. The algorithm employed for the perception of equivalent nodes in a chemical graph has already been reported as part of the isomorphic check in the structure enumeration, which is called the connectivity stack method [5].

The connectivity stack is a sequence of elements in the adjacency matrix describing a chemical structure. When the elements in the upper triangle of the matrix are designated as  $a_{ij}$ , the stack  $(a_1, a_2, a_3, \dots, a_k, \dots)$  corresponds to only one matrix (see Fig. 1), where

$$k = i + (j - 2)(j - 1)/2 \quad (i < j).$$

Thus, the stack represents a single structure unambiguously.

As structures created by the structure generator in the system are made canonical, the following procedure and rules can be applied to all candidate structures. First, nodes of the same kind in the graph are considered. In Fig. 2 (a), for example, A, B and C are methylene groups (imagine methylcyclobutane). The type of node, e.g., the methylene group, is called a component and is already known at this stage of a spectral analysis. Any pair of components of the same kind is permuted in the graph as shown in Fig. 2 (b, c). The permuted graphs are also made canonical, and when the canonical stack of the permuted graph is equivalent to the original stack, the two nodes used for the permutation are recognized as equivalent to each other. Thus the structure shown in Fig. 2 (a) consists of only four topologically

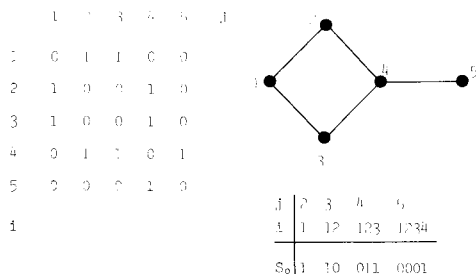


Fig. 1. Adjacency matrix and connectivity stack  $S_0$  for methylcyclobutane.



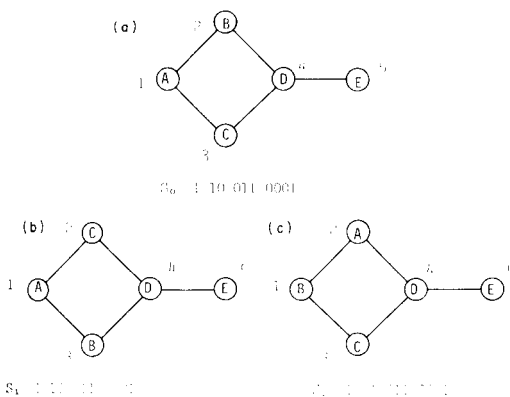


Fig. 2. Procedure for finding equivalent nodes.

distinct carbon atoms because node B is found to be equivalent to node C, i.e., stack  $S_1$  (but not stack  $S_2$ ) is equal to the original stack  $S_0$ . When the permutations with respect to such a pair of nodes have been exhaustively assessed, the perception of the topological symmetry is completed in the proposed method.

## EXPERIMENTAL

A collection of 146 compounds mainly selected from the API 44  $^{13}\text{C}$ -n.m.r. spectral data and the Johnson—Jankowski compilations [6], were used to assess the performance of the  $^{13}\text{C}$ -n.m.r. prediction. The data set consisted of the observed number of signals and the connection table as used normally in the system. The structure types of the compounds in the data set are shown in Table 1, and the prediction results for these compounds are summarized in Table 2.

TABLE 1

Structure types of compounds used for testing the prediction ability

Structure type	No. of compounds	Structure type	No. of compounds
Acids	4	Hydrocarbons	36
Alcohols	13	Ketones	14
Aldehydes	3	Lactone	1
Aromatic compounds (functionalized)	19	Phenols	2
Multifunctional monocyclic compounds	16	Quinones	2
Multifunctional polycyclic compounds	19	Steroids	3
Esters	7	Saccharide	1
Esthers	6		
Total no. of compounds			146

TABLE 2

Results of prediction for 146 compounds by the unmodified program (A) and by the EXAMINE C13 program (B)

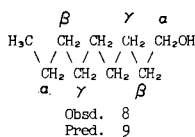
$\Delta$ value <sup>a</sup>	<-3	-3	-2	-1	0	1	2	3	>3
A, No. of compounds	0	1	1	21	93	20	4	2	4
B, No. of compounds	0	1	1	9	110	19	5	1	0

<sup>a</sup>The difference between the predicted number of signals and the observed number of peaks

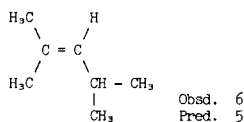
As shown in Table 2, the prediction was completely satisfactory for 93 compounds. The 53 compounds predicted erroneously generally had the following substructures in their molecules: a long alkyl chain, an isopropylidene group, or a *geminal* dimethyl with a chiral center in the ring system, as shown in Fig. 3. The reasons for failure are that all the methylenes in the aliphatic chains are topologically distinct, but not all the signals are separated. Further, methyl groups in the above-mentioned environments are topologically equivalent, but they are different from a stereochemical point of view.

In order to predict more precisely the number of signals from such compounds, the predicting program was modified as follows. In the case of compounds with long alkyl chains (e.g., Fig. 3a), only the methylenic carbons at the  $\alpha$ ,  $\beta$ , and  $\gamma$ -positions to a methyl or hydroxyl group are assumed to appear as individual signals with different chemical shifts, and all other methylenes are assumed to be at the same position. Thus the prediction becomes much more powerful for compounds with more than seven methylene groups. For compounds with isopropylidene groups (Fig. 3b), the number of

(a) 1-Nonanol



(b) 2,4-Dimethyl-2-heptene



(c) 3,3,4-Trimethylcyclohexanone

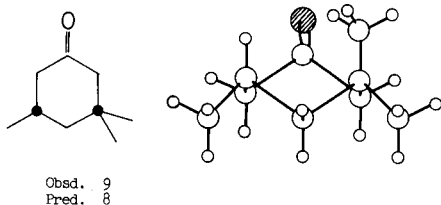


Fig. 3. Examples of compounds predicted erroneously.

TABLE 3

Results for several compounds computed by CHEMICS and with the aid of the EXAMINE module. The error range<sup>a</sup> used is indicated in parentheses

Compounds	Molecular formula	Number of candidates	
		CHEMICS	CHEMICS with EXAMINE C13
Indan	C <sub>9</sub> H <sub>10</sub>	4	1 (0)
Benzil	C <sub>14</sub> H <sub>10</sub> O <sub>2</sub>	20	1 (±1)
Azulene	C <sub>10</sub> H <sub>8</sub>	18	5 (0)
Camphor	C <sub>10</sub> H <sub>16</sub> O	32	26 (0)
2-Cyclohexylcyclohexanone	C <sub>12</sub> H <sub>20</sub> O	147	63 (±1)
2,3-Dimethylnaphthalene	C <sub>12</sub> H <sub>12</sub>	273	19 (0)

<sup>a</sup>Allowance for the difference between predicted signal number and observed number of peaks.

isopropylidene groups present in the molecule is added to the predicted number of signals. However, when an isopropylidene group is surrounded by equivalent carbons as, e.g., in isopropylidene cyclohexane, no addition is necessary because the two methyls are equivalent. For compounds with *geminal* dimethyl groups and a chiral center in a ring system, the program [7] for finding the smallest set of smallest rings is first applied to the adjacency matrix and then, if the *geminal* dimethyl is no farther than the fifth carbon from the chiral center, the two methyl groups are considered to be nonequivalent.

The modified program is called EXAMINE C13. The prediction results given by this program for the compounds listed in Table 1 are summarized in Table 2. The prediction ability is significantly improved by the use of the modified program.

Although the prediction is still erroneous for some compounds, an "error range" can be identified (see Table 3) which the user may take into account as he likes in removing redundant candidates to make the whole system more practical and pragmatic.

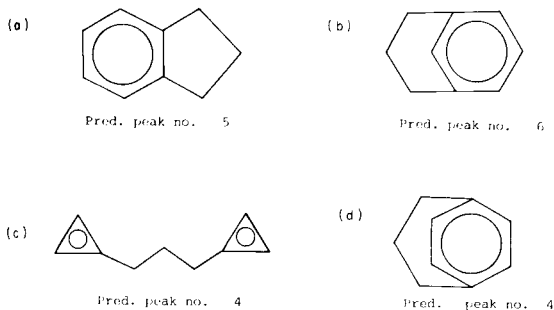


Fig. 4. Structures given by CHEMICS from <sup>1</sup>H- and <sup>13</sup>C-n.m.r. of indan. Numerals 5, 6, 4 and 4 are the predicted numbers of signals for the different structures.

```

*COMMAND?
EXAMINE C13

ERROR RANGE?
0

PREDICTED PEAK?
NO

```

SAMPLE.NO.26/80.OCT.8

\*\* TOTAL NUMBER OF STRUCTURES = 4

PEAK PREDICTION RESULTS

# OBSERVED SIGNAL NUMBER 5

PREDICTED NO. OF STRUCTURES

SIGNAL NO.

5 1

\*\* TOTAL NO. OF HITTING STRUCTURES = 1

\$\$\$ END OF PREDICTION \$\$\$

Fig. 5. Prediction of  $^{13}\text{C}$ -n.m.r. signals for one (a, Fig. 4) of four compounds generated by CHEMICS.

The program is written in FORTRAN IV, and the CPU time required for this experiment was approximately 20 ms, for most compounds, run on a large computer (HITAC M-200H).

## RESULTS

The program EXAMINE C13 which predicts the number of signals expected in  $^{13}\text{C}$ -n.m.r. spectra was introduced to eliminate inappropriate

TABLE 4

$^1\text{H}$ - and  $^{13}\text{C}$ -n.m.r. of indan (input data for CHEMICS)

Molecular formula:  $\text{C}_9\text{H}_{10}$

$^1\text{H}$ -n.m.r. data			$^{13}\text{C}$ -n.m.r. data			
No.	Position (Hz)	Area	No.	Position (ppm)	Height	Multiplicity
1	434.0	81	1	25.5	2319	3
2	429.8	622	2	32.9	5340	3
3	425.8	69	3	124.3	4972	2
4	181.9	150	4	126.1	5558	2
5	175.0	448	5	143.8	1788	1
6	167.8	274				
7	136.8	41				
8	129.8	110				
9	122.0	142				
10	115.0	89				
11	107.5	42				

structures generated for unknown compounds by CHEMICS, by better use of the spectral data. The program was connected to the exit of the CHEMICS system, and test runs for indan and some other compounds were carried out. For example,  $^1\text{H}$ -n.m.r. and  $^{13}\text{C}$ -n.m.r. spectral data of indan (Table 3) were sent to the system which suggested four candidates (Fig. 4) consistent with the input data. The prediction of the number of signals was produced for each of those four structures, immediately after their generation. The computer output (cf. Fig. 5) identified only the first structure (a in Fig. 4) as having five signals, and thus as the most plausible structure. Results for other compounds are listed in Table 3, where "error range" is used for two samples.

The authors thank the Computer Center, Institute for Molecular Science, for the use of the HITAC M-200H computer and for financial support by the Ministry of Education through the fund for "Trace Characterization". We also acknowledge with gratitude the ring-finding program provided by B. Schmidt and J. Fleischhauer, RWTH Aachen, West Germany.

#### REFERENCES

- 1 S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito and T. Yamasaki, *J. Chem. Inf. Comput. Sci.*, 18 (1978) 211.
- 2 S. Sasaki, I. Fujiwara, H. Abe and T. Yamasaki, *Anal. Chim. Acta*, 122 (1980) 87.
- 3 C. A. Shelley and M. E. Munk, *Anal. Chem.*, 50 (1978) 1522.
- 4 C. A. Shelley and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 17 (1977) 110.
- 5 Y. Kudo and S. Sasaki, *J. Chem. Doc.*, 14 (1974) 200; *J. Chem. Inf. Comput. Sci.*, 16 (1976) 43.
- 6 American Petroleum Institute Research Project 44, Selected  $^{13}\text{C}$  Nuclear Magnetic Resonance Spectral Data, Thermodynamics Research Center, Texas, 1975. L. F. Johnson and W. C. Jankowski, *Carbon-13 NMR Spectra*, Wiley-Interscience, New York, 1972.
- 7 B. Schmidt and J. Fleischhauer, *J. Chem. Inf. Comput. Sci.*, 18 (1978) 204.

## DEVELOPMENT OF A NEW FILE SEARCH SYSTEM FOR NUCLEAR MAGNETIC RESONANCE SPECTRA

### Production of an Enlarged Data Base and Search Test

YUZURU KATAGIRI<sup>\*a</sup>, KENZO KANOHTA<sup>b</sup>, KAZUHIKO NAGASAWA<sup>c</sup>,  
TADAO OKUSA<sup>d</sup>, TOSHIO SAKAI<sup>e</sup>, OSAMU TSUMURA<sup>f</sup> and YASUHIKO YOTSUI<sup>g</sup>

<sup>a</sup>*Mitsubishi Rayon Co., Ltd., 3-19 Kyobashi 2-chome, Chuo-ku, Tokyo 104 (Japan)*

<sup>b</sup>*National Institute of Hygienic Sciences, 1-18-1 Kamiyoga, Setagaya-ku, Tokyo 158 (Japan)*

<sup>c</sup>*Toyota Central Research & Development Laboratories, Inc., 41-1 Aza Yokomichi Oaza Nagakute, Nagakute-cho, Aichi-gun, Aichi 480-11 (Japan)*

<sup>d</sup>*Chisso Petrochemical Corporation, 5-1 Goikaigan, Ichihara-shi 290 (Japan)*

<sup>e</sup>*Central Research Laboratory, Mitsubishi Petrochemical Co., Ltd., 1315 Wakaguri, Ami-machi, Inashiki-gun, Ibaraki 300 (Japan)*

<sup>f</sup>*Idemitsu Kosan Co., Ltd., 1280 Kamiizumi, Sodegaura-machi, Kimitsu-gun, Chiba 292-01 (Japan)*

<sup>g</sup>*Daiichi Seiyaku Co., Ltd., 16-13 Kitakasai 1-chome, Edogawa-ku, Tokyo 132 (Japan)*

(Received 23rd January 1981)

### SUMMARY

A practical search system for proton n.m.r. spectra is reported. The coding rules and search algorithms are described in detail. Data for 8000 spectra have been converted into a computer-readable file from printed charts. Several search tests are used to evaluate the usefulness of the search system, and various effects of experimental conditions such as different instruments, frequencies and solvents on recall efficiency are described. The results presented indicate that the system should be applicable to routine analytical work.

Several useful information retrieval systems for infrared, mass and <sup>13</sup>C-n.m.r. spectra have been developed so far, but not as yet for <sup>1</sup>H-n.m.r. (p.m.r.) spectra [1–4]. One of the authors has established the basis of a new file search method for p.m.r. spectra, and reported the relevant coding rules and search algorithms previously [5, 6].

In order to apply the new method to practical use, the present authors have organized a research consortium, and have started the work of producing a complete data base using the world-known Sadtler collection, studying more sophisticated search algorithms since October 1979. The structures of the previous system [5, 6] were re-examined in detail, and new specifications were prepared by adding another two input items to the previous one. So far, 8000 Sadtler spectra have been encoded into a

computer-readable file in the revised format. Moreover, model search tests (cross-check) have been made between corresponding standard data, and several experiments have been conducted by using spectra measured specially for evaluation of the practical usefulness of the system.

The main features of this search system and experimental results are described here.

Recently, an interactive search system for p.m.r. spectra was reported by Bremser [7]. His coding method based on spectral multiplicity (from spin-spin coupling) and proton number (integration of peak intensities) is partly similar to the present coding rules, but the input method of absorption bands (chemical shifts) and the search algorithm seem to be fundamentally different.

## FEATURES OF THE SEARCH SYSTEM

### *Concept of the system design*

Computer-assisted file searching and data bases of p.m.r. spectra have not yet been developed on the practical level, probably because of the difficulties encountered in digitizing these spectra. Spectra from p.m.r. have complicated features, significantly different from  $^{13}\text{C}$ -n.m.r., infrared and mass spectra. Firstly, the multiplicity caused by spin-spin coupling is very important, as well as chemical shift and intensity. Secondly, changes in spectral pattern are often caused by differing experimental conditions (solvent, concentration and temperature); in particular, signals of polar functional groups shift their locations easily and change their shapes. Thirdly, the frequency and actual resolution of instruments used also affect minute spectral patterns. Accordingly, it would be insufficient for the final purpose, even to prepare a precise file which consists only of digitized position and intensity of each peak. The development of a useful data base of p.m.r. spectra seems to be possible only when the features mentioned above are introduced into both coding rules and search algorithms.

### *Coding rules*

First of all, the coding rules proposed by Katagiri [6] were re-examined and the following principles were adopted for the coding rules: (1) eight code names are used to identify signal patterns in a given spectrum as shown in Table 1; (2) values of the chemical shift of absorption bands (except singlets) are encoded as, e.g., 3.52–4.05 ppm whereas, for a singlet, the center position is digitized as 2.10 ppm. In this case, the multiplicity code names shown as D, T, Q and QN are used in a broad sense including asymmetrical spectral patterns to be estimated as apparently first-order, as well as generally symmetrical patterns.

TABLE 1

## Definitions of code names used

Code name	Definition
A	Necessary signal (This may also mean that the signal is a complex multiplet unless the multiplicity specified.)
U	Signal of proton attached to polar functional groups such as -OH, -COOH, etc.
S,D,T,Q	Singlet, Doublet, Triplet, Quartet, respectively
QN	Quintet including sextet, septet, etc.
H	The highest peak in a given spectrum.

*Contents of the new data base*

Details of the data file in process of production are given in Table 2. In the new data file, two more items (molecular formula and proton number corresponding to signal intensities) were added to improve the usefulness of the previous file which contained only the minimum items required for searching. The proton number was added to the data base in order to supplement the absence of quantitative information for signal intensity in this system. Then the new system (PROGRESS; proton magnetic resonance search system) will additionally allow search via molecular formula, the center of gravity and the standard deviation of spectral pattern, for more accurate selection. An input example of a typical standard spectrum is shown in Fig. 1. The data length for one record (or spectrum) is 450 bytes including preparatory bytes.

*Generation of the search file*

The data base prepared by the coding rules described above was changed to a data structure suitable to an automatic searching. The conversion procedures were as follows: (1) compression of the data length from 450 to 150 bytes; (2) special treatment of U code signal; (3) calculation of "KEY DATA"; and (4) calculation of the center of gravity and the standard deviation in a spectrum. The key data show the main characteristics of a given spectrum, and contribute to the improvement of efficiency in the data search.

TABLE 2

## Revised contents of the data base under development

- (1) REGISTER No.
- (2) SERIAL No. & PUBLISHER'S NAME
- (3) MOLECULAR FORMULA<sup>a</sup>
- (4) CODE NAMES
- (5) No. OF PROTON<sup>b</sup>
- (6) CHEMICAL SHIFT

<sup>a</sup>Revised. <sup>b</sup>New entry added.



REGISTRY NO. 20919 SERIAL NO. 20919M PUBLISHER'S NAME SA  
 2-ETHOXY-2-PHENYLACETOPHENONE C<sub>16</sub>H<sub>16</sub>O<sub>2</sub>

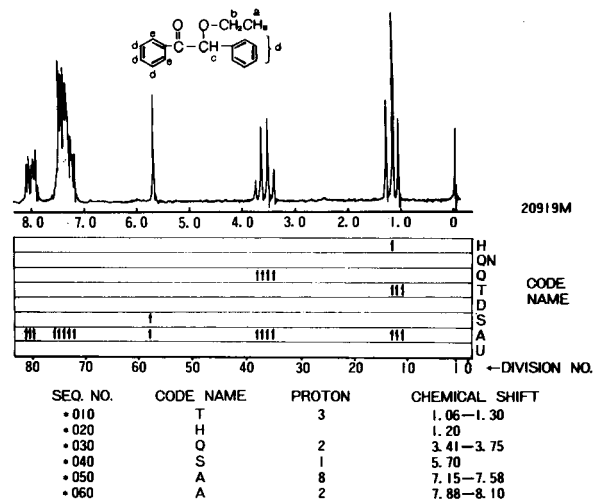


Fig. 1. An example of a Sadtler spectrum (20919M) coded according to the coding rules. The upper part is the spectrum of 2-ethoxy-2-phenylacetophenone. The lower part shows the actual input data for this spectrum. The middle part shows a spectral image pattern obtained from the search master file, showing how digital data are processed in the search system. Note that the spacing of the division number on the abscissa is 0.10 ppm wide.

### Search algorithms

The search algorithms may be outlined as follows. First, logical AND and exclusive OR operations are done about the key data between an unknown spectrum and standard data. Operations for the key data are summarized in Table 3. Secondly, the value of the center of gravity in the spectrum is checked to find whether or not it lies in the permitted range (optional step). Thirdly, code names and the existence of signals are compared between the unknown data and those standard data which passed the first and second check routine. When no match is found between the two sets of data, definite factors are subtracted from initial scores (e.g. 100) of the standard data. Finally, the standard data which have residual scores larger than or equal to a threshold level (usually 50—80 scores) are printed out as hit data. An example of the output list for hit data is illustrated in Fig. 2.

TABLE 3

Contents of key data and operations for them  
 (X): Unknown data. (Y): Standard data

Key data	Operations	Conditions
(a) Element (Presence)	AND	(X) AND (Y) = (X)
(b) Highest peak position	AND	(X) AND (Y) ≠ 0
(c) No signal band	EXOR	(X) EXOR (Y) ≤ TLVAL1

```

*** Q. NO. = 1 *** HIT COUNT = 6 (TOTAL = 7999) ***
COND. : TLVAL1 = 1 , TLVAL2 = 60 , TLVAL3 = - , ELEM = -
      PASS : ELEMENT      = 7999
            H-PEAK       = 1862
            NOT-SIGNAL   = 288
            GRAVITY      = 288
            SIGNAL-CD    = 6

SCORE ID.NAME NOTE GRAVITY ELEMENT SCORE ID.NAME NOTE GRAVITY ELEMENT
  72  27307M SA  8.01  O S      66  25876M SA  7.79  O
  64  26496M SA  7.90  N      64  26572M SA  8.10  N O CL
  64  27317M SA  8.08  N S      60  25883M SA  8.00  O BR

```

Fig. 2. An example of the output list of hit data. The numbers in the "PASS" heading denote numbers of hits at each search step. Here six hits were extracted finally from the enlarged data base, and their scores, serial number (ID. NAME), publisher's name, the center of gravity, and elements were typed out.

The four kinds of search conditions, except for the designation of elements, are called TLVAL1, TLVAL2, TLVAL3 and FC, respectively. TLVAL1 is the tolerance for the result of logical operation of one of the key data. TLVAL2 is a minimum score of hit data. TLVAL3 shows the tolerance for the center of gravity. FC means a set of minus scores indicating disagreement of code names and signal positions (chemical shifts). A set of standard values for FC settled in the search program in advance is generally used if the set is not explicitly designated.

Computer programs for constructing the data base, generating the search file and conducting the main search were written in PL/I. The main search program was also written in FORTRAN. They were processed as an off-line job on an IBM 370/158 system.

#### PERFORMANCE EVALUATION OF THE SEARCH SYSTEM

The performance of the p.m.r. spectral search system was evaluated as follows. First, when search conditions for a given spectrum are fixed, the number of hit data including noise data should depend on the number of data in the data base. The number of hit data can be controlled arbitrarily as desired by simple changes of search conditions, independently from increasing of the data. Secondly, the experimental conditions for the p.m.r. spectra should not have any serious effects on the search efficiency.

The cross-check tests were conducted as a step for the first verification. The second condition is very important in the case of p.m.r. spectra, because the spectra will be changed by frequency, solvent and temperature. In practice, various frequencies, e.g., 60, 80, 90, 100, 200 and 250 MHz, are universally used. For verification of the appropriateness of the coding rules and algorithms, actual search tests were conducted for several measured spectra.

#### *Cross-check tests*

Model search tests, under the same search conditions, were conducted on each spectrum in the data base, considering it as an unknown, when the data

base amounted to 2000 entries (Sadtler serial nos. 25001M–27000M). The results of these tests are shown in Fig. 3 along with those of the previous 500 entries (9501M–10000M). The greater the data base, the more noisy data appeared. But the distribution pattern of the recall efficiency seemed to be quite similar between 500 and 2000 data and was not heavily dependent on the volume of the data base. In addition, it was ascertained that the number of hits could be controlled by changing the search conditions (TLVAL1, TLVAL2 and TLVAL3).

### Effects of experimental conditions

The p.m.r. spectra of several compounds were recorded in different laboratories under arbitrary conditions, and the effects of instruments, frequencies and solvents on the recall efficiency were studied.

The spectra of quinizarin measured with four different instruments under the same measurement conditions were compared. These spectra are shown in Fig. 4. Variations in resolution were observed among the instruments used, but changes in chemical shift (band position in ppm) were not very significant. The results of the search tests (Table 4) indicate that the effects of different instruments and frequencies have little significance in these cases.

The spectra of phthalide measured for three solvents are shown in Fig. 5. The solvent effects on the spectral patterns are not serious, and there was virtually no influence on the recall efficiency (Table 5). In the case of 2-chloroanthraquinone (spectra not shown) differences in solvent effects

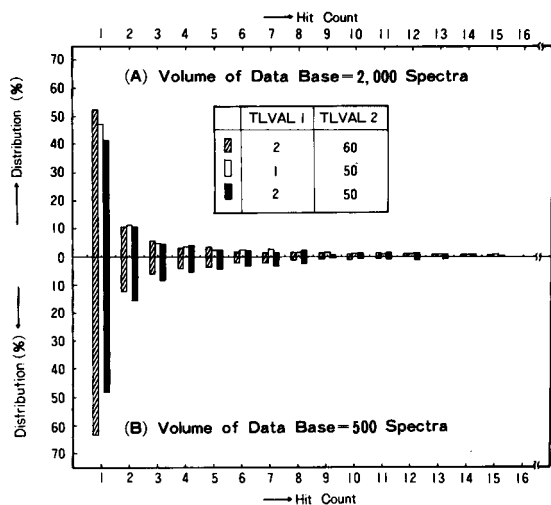


Fig. 3. Results of the cross-check tests for the enlarged data base (A) and the small data base (B). Distribution (%) means how many times a given number of hits were obtained through the tests among all the data.

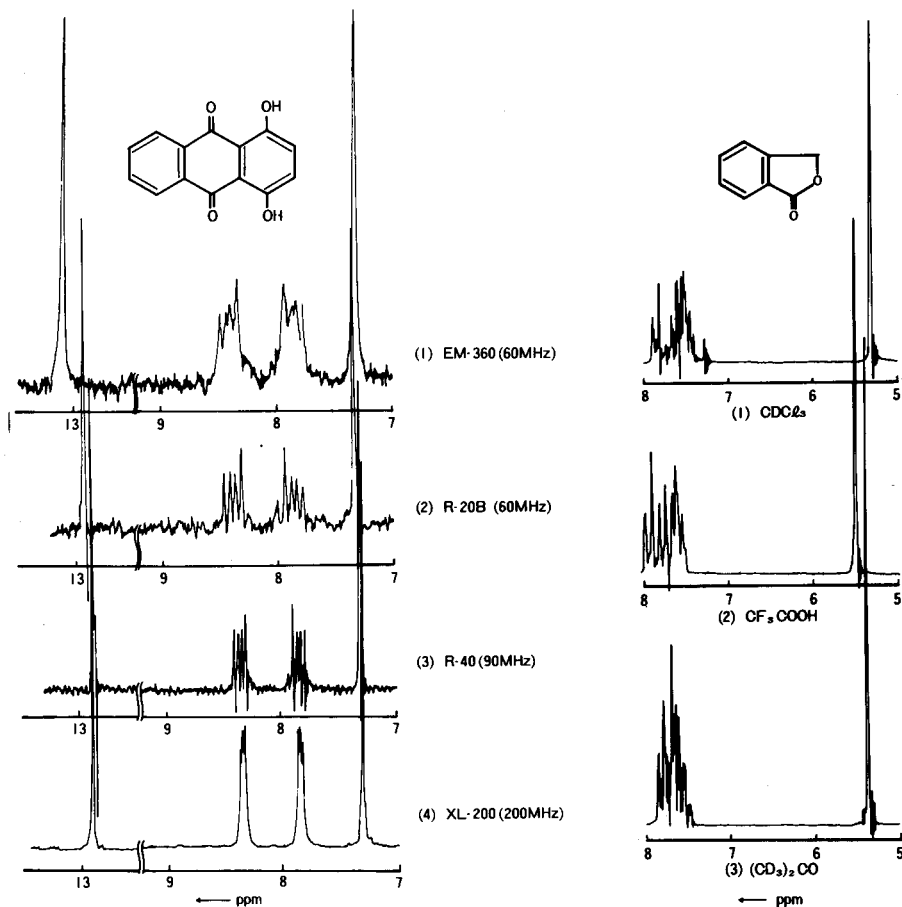


Fig. 4. Spectra of quinizarin (1,4-dihydroxyanthraquinone) measured with four different spectrometers.

Fig. 5. Spectra of phthalide ( $C_8H_6O_2$ ) measured in various solvents with the same spectrometer.

were obvious between  $CDCl_3$  and  $DMSO-d_6$  and the correct answer could not be obtained.

#### *Quantitative effects of the size of the data base*

Search tests were applied for the various spectra according to the increasing size of the data base. These spectra and the results of the search tests are given in Figs. 6 and 7. The spectra of phenanthrenequinone has signals only in a narrow low-field range (7–8 ppm), but correct answers were extracted with a high degree of efficiency by setting the conditions appropriate to the spectral pattern (I, Fig. 6). *D*-Camphor has signals only in the high-field range (0.8–2.4 ppm) and correct answers were obtained under normal conditions (II, Fig. 6).

TABLE 4

Effects of instrument and frequency used on the recall efficiency<sup>a</sup>  
 (Sample, Quinizarin (C<sub>14</sub>H<sub>8</sub>O<sub>4</sub>); Sadtler No. 25876M; Solvent, CDCl<sub>3</sub>)

Instrument	Freq. (MHz)	Number of hits					Score and ranking of correct answer
		DB = 2500	DB = 3500	DB = 5500	DB = 6500	DB = 8000	
EM-360	60	6	6	6	6	6	66 (2/8000)
R-20B	60	8	8	10	11	11	78 (2/8000)
R-40	90	7	8	10	11	11	62 (3/8000)
XL-200	200	7	8	10	12	12	68 (1/8000)

<sup>a</sup>TLVAL1 = 1, TLVAL2 = 60, TLVAL3 = --, ELEM = --.

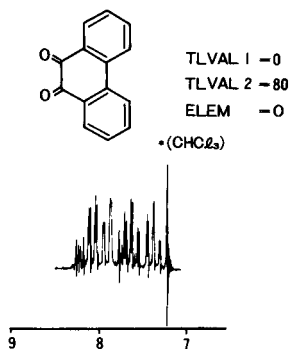
TABLE 5

Effects of solvents used on the recall efficiency for Sadtler no. 27216M (C<sub>8</sub>H<sub>8</sub>O<sub>2</sub>)<sup>a</sup>

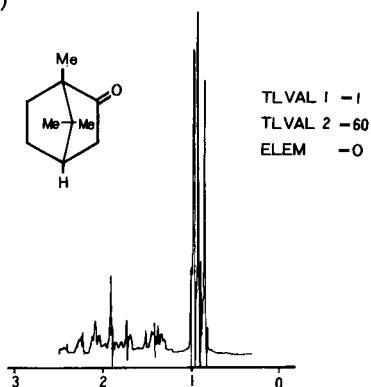
Solvent	Number of hits						Score and ranking of correct answer
	DB = 2500	DB = 3500	DB = 4000	DB = 5500	DB = 6500	DB = 8000	
CDCl <sub>3</sub>	6	8	8	9	10	11	90 (1/8000)
CF <sub>3</sub> COOH	4	4	4	5	5	7	64 (5/8000)
(CD <sub>3</sub> ) <sub>2</sub> CO	5	7	7	7	7	8	90 (1/8000)

<sup>a</sup>TLVAL = 1, TLVAL2 = 50, TLVAL3 = --, ELEM = O.

I)



II)



SAMPLE	FREQ. (MHz)	NUMBER OF HITS					SA NO.	SCORE	RANKING	SEARCH CONDITIONS		
		DB-2500	4000	5500	6500	8000				TLVAL 1	TLVAL 2	ELEM
I	100	3	4	5	7	9	27364M	80	4/9	0	80	0
II	100	2	3	4	4	7	27185M	76	1/7	1	60	0

Fig. 6. Spectra and results of search tests for phenanthrenequinone(I) and D-camphor(II).

III)

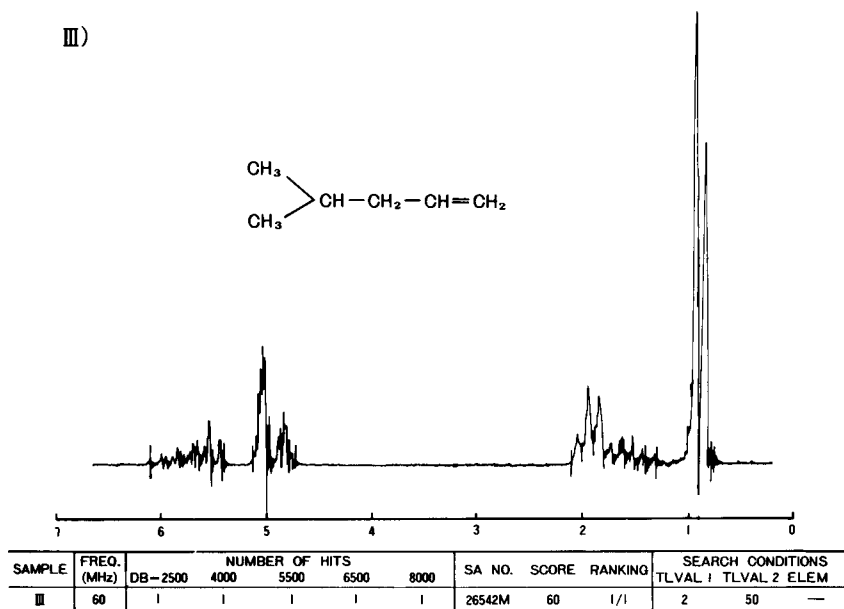


Fig. 7. Spectrum and results of search tests for 4-methyl-1-pentene.

For 4-methyl-1-pentene, a correct answer with 100% recall efficiency (hit = 1) was obtained under normal search condition, regardless of the size of the data base (Fig. 7).

These results strongly suggest that the search system should be applicable in routine analytical work. The data base is planned to contain some 30,000 entries within a year.

We express our sincere gratitude to Mr. T. Matsukura, Sanyo Shuppan Boeki Co., Inc. (the consortium office) for his coordinating efforts with Sadtler Research Laboratories, Inc.

#### REFERENCES

- 1 S. R. Heller and R. J. Feldmann, *J. Chem. Educ.*, 49 (1972) 291.
- 2 S. Sasaki, Y. Yotsui and S. Ochiai, *Japan Analyst (Bunseki Kagaku)*, 24 (1975) 213.
- 3 S. Sasaki, in S. Fujiwara and H. B. Mark, Jr. (Ed.), *Information Chemistry*, The University of Tokyo Press, Tokyo, 1975, p. 227.
- 4 O. Yamamoto and M. Yanagisawa, 18th NMR Symposium, Osaka, Japan, Nov. 4-6, 1979.
- 5 Y. Katagiri, 20th FEChem Conf. on Computer-Based Analytical Chemistry, Portorož, September, 1979, p. 69.
- 6 Y. Katagiri, Japanese Published Unexamined Patent Application No. 141190/79.
- 7 W. Bremser, *Chem. Zeit.*, 104 (1980) 53.

## GENERATING RULES FOR PAIRS — A COMPUTERIZED INFRARED SPECTRAL INTERPRETER

HUGH B. WOODRUFF\* and GRAHAM M. SMITH

*Merck Sharp & Dohme Research Laboratories P.O. Box 2000, Rahway, New Jersey 07065 (U.S.A.)*

(Received 23rd January 1981)

### SUMMARY

A computerized spectral interpreter that attempts to parallel the reasoning used by a spectroscopist can only be as effective as the rules it employs. A detailed description of the process by which these interpretation rules are generated in the form of binary decision trees is presented. The aldehyde functionality is selected for the demonstration. Through the use of an English-like computer language to express the rules, the two major goals for the computerized interpreter of understandability and flexibility are achieved.

Accurate interpretation of infrared spectra with the aid of a computer is a goal sought in a variety of laboratories [1–12]. While many of the algorithms produced through these investigations can be applied successfully to only a selective group of chemical functionalities, several computer programs exist which show considerable promise in interpreting infrared spectra of compounds containing a large diversity of functional groups [13–18]. Work recently reported from this laboratory [19] describes PAIRS (Program for the Analysis of IR Spectra) and demonstrates its utility on several sample spectra. PAIRS is available for distribution from the Quantum Chemistry Program exchange, Bloomington, Indiana 47405 (catalog no. QCPE 426).

### DESCRIPTION OF PAIRS

PAIRS is designed to parallel as closely as possible the reasoning a chemist uses in interpreting infrared spectra. Two paramount considerations during the development of PAIRS were to make all aspects of the computerized interpreter readily understandable to the spectroscopist and to maintain a high degree of system flexibility. If necessary, computer efficiency would have been sacrificed in order to simplify the chemist/computer interface. A detailed description of PAIRS and the manner in which understandability and flexibility are maintained has been presented [19], so only a cursory description of the system components is presented here.

PAIRS utilizes information from three different sources: (1) the spectrum to be interpreted, digitized in a manner to produce peak locations between 4000 and 500  $\text{cm}^{-1}$ , peak intensities ranging from 1 (very weak) to 10 (very

strong), and peak widths of 1 (sharp), 2 (average), or 3 (broad); (2) supplemental information such as sample state and molecular formula; and (3) a set of interpretation rules. Obviously, the legitimacy of these rules is the key factor in determining the accuracy of the interpretation. Faulty interpretations can be improved upon by modifying or augmenting the existing collection of rules, and this correction process ideally should be implemented by the spectroscopist. The design of PAIRS makes the scientist much better able to cope with this assignment, since the rules are treated as data by the interpreter and need not be hard-coded in a computer language that the chemist has difficulty in understanding. Rather, the rules are expressed in an English-like language called CONCISE (Computer Oriented Notation Concerning Infrared Spectral Evaluation). Anytime a scientist disagrees with an interpretation made by PAIRS, these CONCISE rules can readily be used to determine how PAIRS reached its conclusions and the offending rule(s) may be altered to suit the individual.

While the earlier general description of PAIRS [19] does discuss the CONCISE language and demonstrate its utility with a sample decision tree, a more detailed description of the process by which a decision tree is generated should prove useful to spectroscopists attempting to improve upon the current set of rules. Such a detailed description is presented in this paper with the aldehyde functionality selected for the demonstration.

#### *Generating a decision tree*

While the obvious initial step in the overall procedure of generating interpretation rules is to create the first tree, the subsequent improvements of the rules should be an ongoing process (see Fig. 1). Initially, information on interpreting aldehydes must be obtained from reference books [20–26], journal articles [27], or any other available sources. The scientist accumulates this information and decides how spectra should be interpreted for the presence of an aldehyde group.

Now a critical point in generating decision trees and thereby in determining the effectiveness of PAIRS has been reached. The scientist must express these accumulated facts as clear and accurate rules. For example, the consensus of the reference sources is that for aldehydes: (1) a strong carbonyl peak will appear between 1765 and 1660  $\text{cm}^{-1}$ , the exact position depending on the environment near the carbonyl group; (2) two moderately intense bands will appear between 2900 and 2695  $\text{cm}^{-1}$ , usually near 2820 and 2720  $\text{cm}^{-1}$ . Since the 2820  $\text{cm}^{-1}$  peak may sometimes overlap with C–H stretching bands originating from other parts of the molecule while the 2720  $\text{cm}^{-1}$  peak is usually sharp and clearly separated from other C–H stretching bands, the end result is that “medium intense absorption near 2720  $\text{cm}^{-1}$  (3.68  $\mu\text{m}$ ) accompanied by a carbonyl absorption band is good evidence for the presence of aldehyde group” [25].

Converting this accumulated information into interpretation rules in the form of binary decision trees is not a difficult task, but it does require that



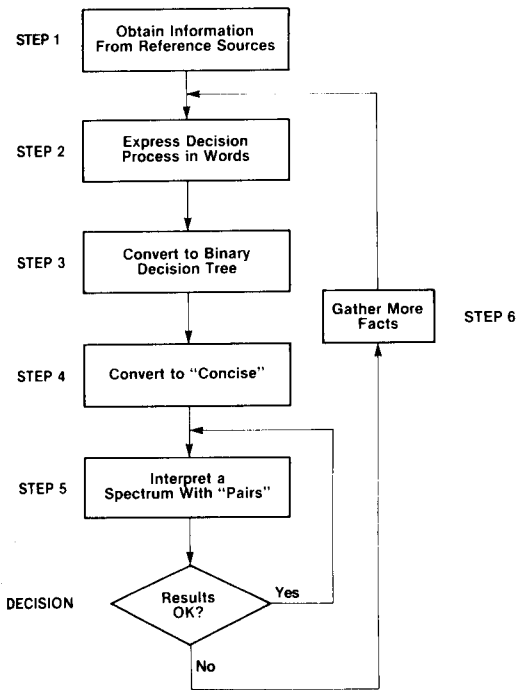


Fig. 1. Generalized process for generating interpretation rules, purposely pictured with no exit point as accumulating knowledge is an ongoing process.

some additional judgments be made. Based on the information described above, an approach is shown in Fig. 2. The mode of presentation is somewhat different from a conventional binary decision tree; however, the alterations result in a more compact display [19]. Following any decision point in a binary tree,

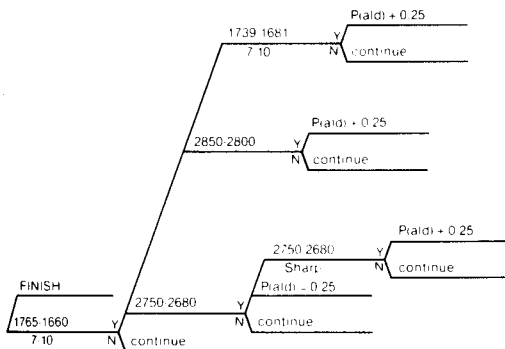


Fig. 2. Initial decision tree for determining probability of aldehyde presence; "continue" indicates the information flow returns to the previous branch point and proceeds to the next command or query.

either the "yes" or the "no" path is followed. Additional queries, probability adjustments, etc., occur in either path, but there may reach a point in either pathway following which everything is the same. Rather than these identical sections being shown multiple times, they are indicated by a line emanating from the left side of the appropriate decision point and terminating at the next query or command.

As a simple example, regardless of whether or not a relatively strong (7–10) carbonyl peak appears between 1765 and 1660  $\text{cm}^{-1}$ , either the "Y" or "N" path would ultimately reach identical points, namely at the "FINISH" command. Presence of an appropriate carbonyl peak causes a subsequent query to be posed concerning the 2750–2680  $\text{cm}^{-1}$  region, while absence of the carbonyl absorption causes no additional operation to be performed other than "FINISH"; hence the aldehyde probability would remain at 0. If a peak is found between 2750 and 2680  $\text{cm}^{-1}$ , then the probability is increased and, should that peak be sharp, it is raised again. Regardless of the outcome of the 2750–2680 decision, the 2850–2800  $\text{cm}^{-1}$  region is checked, with peak presence increasing the probability. Finally, no matter what the decisions concerning the C–H stretching region have been, a narrower carbonyl range is checked and an affirmative answer will increase the probability.

Once the third step shown in Fig. 1 has been completed, it is necessary to convert the decision tree into rules expressed in CONCISE, and the end result is presented in Fig. 3. The if-then-else-terminology required by binary decision trees is present in CONCISE as well as the concept of begin-done blocking of statements. To clarify the blocking, the appropriate BEGIN and DONE statements are connected in the figure.

When this tree is used to interpret a variety of aldehyde and non-aldehyde spectra, the results are not very satisfactory. Several non-aldehyde, but carbonyl-containing compounds have spectra with weak absorption between 2850 and 2800  $\text{cm}^{-1}$ , which in conjunction with an appropriately positioned carbonyl peak results in an aldehyde probability of 0.5. Acid spectra often contain a shoulder or weak absorption between 2650 and 2600  $\text{cm}^{-1}$ . A number of acid spectra containing that absorption at a slightly higher frequency than is typically expected result in a 0.5 probability for aldehyde because of affirmative replies to the 2750–2680 and 1739–1681 questions. Frequently, obvious aldehydes result in probabilities no greater than 0.75 since the peak near 2720  $\text{cm}^{-1}$  is encoded as average in width rather than sharp. While these results do show some promise, large improvements would be desirable.

#### *Improving the interpretation rules*

For those spectra mentioned above, the final step of the process pictured in Fig. 1, the "RESULTS OK" query, has resulted in a negative reply, at least for a program such as PAIRS which is designed to handle a diversity of compounds. Thus, it is necessary to express the decision process once again, this time trying to be more discriminating. An approach found to be effective in

```

IF ANY INTENSITY 7 TO 10 SHARP TO BROAD PEAKS ARE IN RANGE 1765 TO 1660
$
THEN BEGIN
IF ANY INTENSITY 1 TO 10 SHARP TO BROAD PEAKS ARE IN RANGE 2750 TO 2680
$
THEN BEGIN
SET ALDEHYDE TO 0.25
IF ANY INTENSITY 1 TO 10 SHARP PEAKS ARE IN RANGE 2750 TO 2680
$
THEN BEGIN
ADD 0.25 TO ALDEHYDE
DONE
DONE
$
IF ANY INTENSITY 1 TO 10 SHARP TO BROAD PEAKS ARE IN RANGE 2850 TO 2800
$
THEN BEGIN
ADD 0.25 TO ALDEHYDE
DONE
$
IF ANY INTENSITY 7 TO 10 SHARP TO BROAD PEAKS ARE IN RANGE 1799 TO 1681
$
THEN BEGIN
ADD 0.25 TO ALDEHYDE
DONE
DONE
$
$
FINISH
$
END

```

Fig. 3. Aldehyde rules expressed in "CONCISE" language.

generating the many decision trees used by PAIRS is to consider each spectral region of importance to the particular functionality being tested and to determine what interferences might be expected from absorptions by other functionalities. For example, absorption between 2750 and 2680  $\text{cm}^{-1}$  may occur in carboxylic acids or compounds containing  $\text{NH}_2^+$  and have no connection to the C—H stretching in aldehydes. Of course, acids show other prominent absorption patterns, especially the broad peak near 3000  $\text{cm}^{-1}$ , and  $\text{NH}_2^+$  also has a readily identifiable broad absorption in the 2750–2500  $\text{cm}^{-1}$  region. Some other alterations to the decision tree are made apparent from the results of the preliminary testing. Restricting the peak width of the 2750–2680  $\text{cm}^{-1}$  peak to sharp is detrimental, as frequently that peak is of average width, depending on the sampling conditions, etc.

Encoding the absorption by mineral oil if the sampling state is a mull is obviously incorrect, but it has sometimes been done. Thus, the decision tree should take into account that the 2850–2800  $\text{cm}^{-1}$  peak from an aldehyde will not be discernible in mineral oil mulls.

These improvements, along with several others, are incorporated into the revised decision tree shown in Fig. 4. The tree is considerably more complex than the initial tree, but interpreting infrared spectra properly is likewise

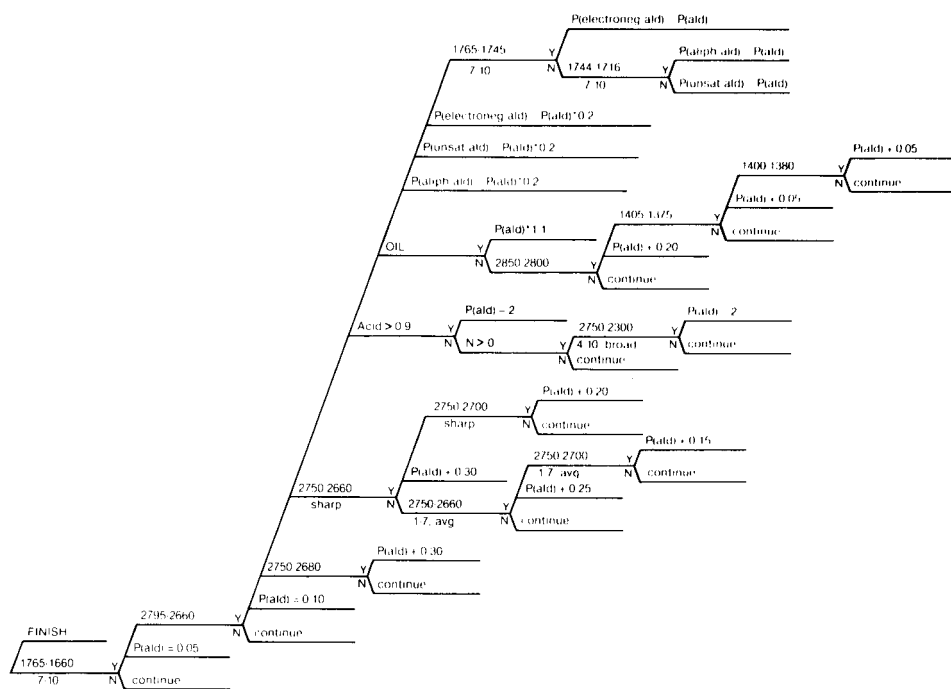


Fig. 4. Modified aldehyde decision tree; see Fig. 2 for description of "continue" statement.

quite complex. The rationale behind most of the changes has already been presented or should be readily apparent; however, several comments are appropriate. In addition to the peak queries illustrated in Figs. 2 and 3, PAIRS allows probability queries, formula queries, and solvent queries. Examples of all query types are included in Fig. 4 ("Acid > 0.9", "N > 0", and "OIL"). A probability query concerning any functionality, aldehyde or non-aldehyde, is permissible at any time; however, a query such as "Acid > 0.9" only makes sense if the acid decision tree has previously been tested by PAIRS.

Rather than continuing to lump all aldehydes together as a single class, three subclasses have been created, saturated (no  $\alpha$ ,  $\beta$  unsaturation), unsaturated (in at least the  $\alpha$ ,  $\beta$  position), and  $\alpha$ -electronegative group. A chemist would attempt to make this distinction, so PAIRS also should do so.

An ideal aldehyde would result in accumulated evidence surpassing a probability of 1.0. It is common for excess of evidence to be ignored once a decision has been made, so allowing probabilities to surpass 1.0 creates no problem, as long as PAIRS makes certain that the final probability falls between 0 and 1. In fact, PAIRS returns a minimum probability of 0.01 and a maximum of 0.99 to highlight the need for the chemist to consider the results carefully, a need which might be overlooked if the computerized interpreter were to indicate absolute certainty with probabilities of 0.0 or 1.0.

## RESULTS

When this decision tree was used, results of interpretations of test spectra from a variety of sources [22–25, 27–30] were vastly improved. Of course, this improvement should come as no surprise because the decision tree shown in Fig. 4 is a direct outgrowth of considerable testing and accumulating of data using earlier versions of aldehyde decision trees. Just as chemists accumulate knowledge and improve their interpretation abilities, the interpretation abilities of PAIRS will continue to improve as long as spectroscopists convey their newly acquired knowledge to the program in the form of upgraded decision trees.

Aldehyde probabilities distributed as shown in Table 1 result from using PAIRS to interpret over 250 spectra, including 42 from compounds containing an aldehyde group. Of the 220 non-aldehydes tested, about 44% of them contain carbonyl groups. The aldehydes range from simple, monofunctional compounds such as acetaldehyde and propionaldehyde, to the quite complex structures of antibiotics like the spiramycins and leucomycins. As might be expected, the simple aldehydes are interpreted the best; however, results from the 23 antibiotics tested [30] are very encouraging as 16 of them yield a 0.99 probability and two others have a 0.85 probability. In fact, only one antibiotic, leucomycin A<sub>5</sub>, is interpreted poorly (0.05) because of no apparent aldehyde C–H stretching absorption. Of course, the most intense band in the entire spectrum is at the 40% transmission level, so the expected weak absorptions near 2720 and 2820 cm<sup>-1</sup> (only one aldehydic C–H bond in a molecule of 119 atoms) are even less discernible.

The two non-aldehydes with high aldehyde probabilities are 4-methyl-2-pentanone [28] (probability = 0.90) and 3 $\beta$ , 17 $\beta$ -diacetoxy-5 $\alpha$ -androstan-16-ylidenacetic acid [29] (probability = 0.99). Both spectra show weak absorption near 2720 cm<sup>-1</sup>, although the former spectrum is so strongly absorbing that the strongest peaks pegged out at the bottom of the chart. Certainly the overall results are very promising as a wide variety of complex spectra have been interpreted.

TABLE 1

Distribution of aldehyde probabilities for spectra interpreted by PAIRS

Probability range	Aldehydes (42) (%)	Non-aldehydes (220) (%)
0.90–0.99	78.6	0.9
0.70–0.89	7.1	0.5
0.30–0.69	9.5	0.9
0.01–0.29	4.8	97.7

## Conclusions

While the improvements are significant when the current decision tree is used, further refinement is always possible. For example, distinction between aromatic,  $\alpha$ ,  $\beta$ -unsaturated, and  $\alpha$ ,  $\beta$ - $\gamma$ ,  $\delta$ -unsaturated aldehydes should probably be attempted. In the 1950's Pinchas [27] indicated that the position of the  $2720\text{ cm}^{-1}$  region C—H stretching peak was altered in spectra of compounds in which the aldehydic hydrogen could be participating in hydrogen bonding. This type of information could also be incorporated into expanded interpretation rules. By utilizing the approach presented in this paper or a similar approach, a chemist should be able to amend or replace existing decision trees or add new trees for other functionalities, thereby developing rules that parallel normal reasoning more closely. The binary decision tree concept and the CONCISE language combine to make PAIRS a success at achieving its initial goals, understandability and flexibility.

The useful suggestions and encouragement of David D. Saperstein, George V. Downing, Peter Gund, and Bert Singleton contributed to the success of this effort and are gratefully acknowledged.

## REFERENCES

- 1 H. Abe and S. Sasaki, *Sci. Rep. Tohoku Imp. Univ., Ser. 1*, 55 (1972) 63.
- 2 H. M. Bell, *J. Chem. Educ.*, 53 (1976) 26.
- 3 V. A. Dement'ev, *J. Appl. Spectrosc. (USSR)*, 23 (1975) 1232.
- 4 N. A. B. Gray, *Anal. Chem.*, 47 (1975) 2426.
- 5 L. A. Gribov, M. E. Elyashberg and M. M. Raikhshtat, *J. Mol. Struct.*, 53 (1979) 81.
- 6 L. A. Gribov, M. E. Elyashberg and V. V. Serov, *Anal. Chim. Acta*, 95 (1977) 75.
- 7 Z. Hippe and A. Kerste, *Bull. Acad. Pol. Sci., Ser. Sci. Chim.*, 22 (1974) 541.
- 8 L. K. Oliver and R. V. Sweet, *Clin. Chim. Acta*, 72 (1976) 17.
- 9 S. Sasaki, H. Abe, K. Saito and Y. Ishida, *Bull. Chem. Soc. Jpn.*, 51 (1978) 3218.
- 10 V. Tamas, M. Odon, R. Jozsef and H. Pal, *Magy. Kem. Foly.*, 84 (1978) 433.
- 11 S. Ungan, *Middle East Tech. Univ. J. Pure Appl. Sci.*, 8 (1975) 305.
- 12 C. G. A. van Eijk and J. H. van der Maas, *Fresenius Z. Anal. Chem.*, 286 (1977) 80.
- 13 R. E. Anacreon and S. C. Pattacini, *Am. Lab.*, 12(2) (1980) 97.
- 14 M. Ford, H. Carter, P. White, J. Coates, A. Savitsky, S. Geary, A. Muir and R. W. Hannah, *Abstracts of the 1979 Pittsburgh Conference*, Paper No. 235.
- 15 J. P. Coates, R. W. Hannah and D. W. Mayo, *Abstracts of the 1980 Pittsburgh Conference*, Paper No. 573.
- 16 H. B. Woodruff and M. E. Munk, *J. Org. Chem.*, 42 (1977) 1761.
- 17 H. B. Woodruff and M. E. Munk, *Res.,/Dev.*, 28(8) (1977) 34.
- 18 H. B. Woodruff and M. E. Munk, *Anal. Chim. Acta*, 95 (1977) 13.
- 19 H. B. Woodruff and G. M. Smith, *Anal. Chem.*, 52 (1980) 2321.
- 20 L. J. Bellamy, *The Infra-red Spectra of Complex Molecules*, Wiley, New York, 1954.
- 21 L. J. Bellamy, *Advances in Infrared Group Frequencies*, Methuen, London, 1968.
- 22 R. T. Conley, *Infrared Spectroscopy*, 2nd edn., Allyn and Bacon, Boston, 1972.
- 23 K. Nakaniishi, *Infrared Absorption Spectroscopy*, Holden-Day, San Francisco, 1962.
- 24 D. J. Pasto and C. R. Johnson, *Organic Structure Determination*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- 25 R. M. Silverstein and G. C. Bassler, *Spectrometric Identification of Organic Compounds*, 2nd edn., Wiley, New York, 1967.

- 26 R. G. White, *Handbook of Industrial Infrared Analysis*, Plenum Press, New York, 1964.
- 27 S. Pinchas, *Anal. Chem.*, 27 (1955) 2; 29 (1957) 334.
- 28 B. Schrader and W. Meier, *DMS Raman/IR Atlas of Organic Compounds*, Verlag Chemie, Weinheim, Germany, 1974.
- 29 *Coblentz Society Spectra*, Vol. 6, Coblentz Society, Philadelphia, 1969.
- 30 C. Lenzen and L. Delcambe, *Infrared Spectra of Antibiotic Substances Part 11*, International Center of Information on Antibiotics, Liège, Belgium, 1973.

## COMPUTER-AIDED DATA REDUCTION OF ABSOLUTE INFRARED INTENSITY MEASUREMENTS BY TRANSMISSION AND REFLECTION†

R. N. JONES\*

*Department of Chemistry, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo 152 (Japan)*

D. G. CAMERON and T. G. GOPLEN

*Division of Chemistry, National Research Council of Canada, Ottawa, Ontario K1A 0R6 (Canada)*

(Received 23rd January 1981)

### SUMMARY

The profiles and intensities of infrared absorption bands of liquids provide information about the dynamic structure of materials in the liquid state. Such data can also provide secondary reference standards for checking intensity measurements made on both dispersive and Fourier transform infrared spectrophotometers. A method for evaluating the absorption intensities of thin films of liquids is described. The results are expressed in terms of the optical constants ( $n$ ,  $k$ ). The method is based on measurement of the film transmission over the full wavenumber range of the measurement combined with an additional measurement of the attenuated total reflection spectrum over a short range in a region of low absorption. Recent measurements on a series of thirteen common organic liquids are cited as are measurements on free standing polymer films.

This paper is concerned with the use of an off-line computer in conjunction with an infrared spectrophotometer to measure the shapes and intensities of the absorption bands of thin films of liquids and solids. In a broader context, it will also stress the need to understand exactly what are the quantities being measured when sophisticated analytical instruments are used. This is true especially where the measurements are made on commercially engineered “black boxes” with impressive control panels carrying knobs and switches with disarmingly simple names printed on them. It applies with added force where the black box is linked to a dedicated computer with its own packaged data reduction program.

The general educational thesis that will emerge is that when chemical measuring instruments linked to computers are used, it is essential to know (a) the quantity that is actually being measured; (b) what the computer program is really doing to the experimental data.

†Contribution No. 19666, Laboratories of the National Research Council of Canada.

\*Permanent address: Ann Manor, Apt 601, 71 Somerset Street West, Ottawa, Ontario, Canada K2P 2G2.



## INSTRUMENTAL CONSIDERATIONS

*The infrared grating spectrophotometer*

The essential features of a modern infrared ratio-recording grating spectrophotometer are: (a) a source of continuous infrared radiation; (b) an optical system to create two matched radiation beams; (c) a rotating mirror to deflect each beam alternately on to the entrance slit and dispersing grating; (d) a detector with associated optics which alternately measures the energy of each dispersed beam; and (e) an electronic recording system which records and stores or plots the ratio of these intensities at each measured wavelength.

The quantities directly measured are the radiant fluxes in the two dispersed beams ( $P_{\text{reference}}$ ,  $P_{\text{sample}}$ ). The quantity recorded is the ratio of these intensities, the apparent transmission  $T_{\text{app}} = P_{\text{sample}}/P_{\text{ref}}$ . Two other derived quantities are often recorded, the apparent absorbance  $A_{\text{app}} = \log_{10} (P_{\text{ref}}/P_{\text{sample}})$ ,

and the apparent absorption index

$$k_{\text{app}} = (4\pi\bar{\nu}d)^{-1} \log_e (P_{\text{ref}}/P_{\text{sample}})$$

where  $\bar{\nu}$  is the wavenumber ( $\text{cm}^{-1}$ ) and  $d$  is the cell thickness (cm).

Molecular spectroscopists usually express absorption intensities in absorbance units. Those concerned with the optical rather than the molecular properties of the absorbing sample prefer the absorption index because it is the imaginary component of the complex refractive index  $\hat{n} = n + ik$ . The absorption index is closely related to the dielectric constant and other optical quantities important in dispersion theory and classical optics.

*Photometric calibration*

It should not be assumed that the ratio recorded by the detector exactly conforms with the true relative beam intensities or that it is linear over the full intensity range. This can be checked by the insertion of a rotating sector photometer in the sample beam [1, 2]. The photometer will attenuate the beam by a precisely known amount. The sector photometer itself is calibrated, either metrically by measurement of the areas of the sector blades, or optically with white light in the visible region using a photocell which has in turn been calibrated with a standard illuminating lamp. By a combination of both methods the infrared spectrophotometer is calibrated in terms of two of the basic S.I. units: the meter by metrication and the candela by the optical method. The two methods give attenuation values for the rotating sectors agreeing to within  $2 \times 10^{-4}$  T. The accuracy of the sector-calibrated spectrophotometer is about  $\pm 2 \times 10^{-3}$  T.

*Fourier-transform infrared spectrophotometers*

Fourier-transform infrared spectrophotometers (F.t.i.r.) are now displacing instruments with the grating type of dispersion for many purposes [3]. They have many advantages, particularly with respect to greater sensi-

tivity in dealing with weak spectra, but from the point of view of absolute photometric accuracy there are difficulties in calibrating them with sector photometers and alternative methods to link them directly with a basic S.I. unit have not yet been formulated. It would be desirable to have some secondary standards for the intensity calibration. Conveniently this could be some suitable absorption bands of known shape and absolute intensity. This would also be desirable for routine checks on dispersion-type instruments because the sector photometer method is time-consuming and tedious and it complicates the computer-aided data reduction.

#### *Factors determining infrared band shapes in condensed states*

Spectroscopists are also concerned with theoretical problems influencing the contours of non-overlapping infrared bands. These can give useful information about the molecular motions in the liquid state [4, 5]. The band shapes and intensities of solid films give useful information about the dichroic properties of the film when polarized infrared radiation is used and this is of particular interest to polymer chemists [6]. There is also need to improve the accuracy of infrared intensity measurements for ordinary quantitative analysis, particularly where the measurements are to be exchanged between different laboratories. Infrared spectrophotometers have always been characterized by high precision which, however, only means that the errors associated with a given instrument are reproducible.

#### *The optics of the absorption cell*

So far, the spectrophotometer has been considered only as a measuring instrument for comparing the two beam intensities. In practice a sample must be inserted and additional problems then arise. There are reflections at the faces of the cell and interference and reflections within the sample itself. These apply with equal force to spectra measured on dispersion or Fourier-transform spectrophotometers.

In the simplest case it is the custom in analytical chemistry to compensate for these effects by placing a "matched" cell in the reference beam. This is only effective, and then to a limited extent, by measuring the spectrum in a dilute solution in a weakly absorbing solvent. Carbon tetrachloride, carbon disulfide or chloroform are widely used for this purpose. This compensation technique cannot be used for measurements on pure liquids, mixtures of absorbing liquids or solid films because there are no appropriate materials that can be used for compensation.

In these cases the reference cell must be left "open" and the radiation dissipated in the sample cell by absorption, reflection and interference must be computed by classical optical theory. These calculations must be made separately at each recorded data point and this has only become a practical possibility with computer-aided data reduction facilities.

These calculations have been discussed in detail elsewhere [7-13]; here, only the outcome of the computation will be indicated briefly. The computation can be formalized as

$$k_{\text{app}}(\bar{\nu}) = f_1(n_{\text{true}}, k_{\text{true}}, n_{\text{window}}, a, p)_{\bar{\nu}} \cdot f_2(d, \theta, \Phi)$$

where  $n_{\text{true}}$  and  $k_{\text{true}}$  are the real and imaginary components of the true complex refractive index ( $\hat{n}_{\text{true}}$ );  $n_{\text{window}}$  is the refractive index of the (transparent) window material;  $a$  is the photometric correction from the sector photometer measurement; and  $p$  is a correction for the polarization discrimination of the spectrophotometer. These are grouped into  $f_1$  as they are functions of the wavenumber ( $\bar{\nu}$ ). Under  $f_2$  are grouped three parameters that are dependent only on the geometry of the system:  $d$  is the mean thickness of the cell;  $\theta$  is a correction for the non-parallelism of the radiation beam; and  $\Phi$  is a correction for the non-parallelism of the cell windows.

Unfortunately, though this algorithm permits the computation of  $k_{\text{app}}$  if  $k_{\text{true}}$  is known, it cannot be simply transposed algebraically to evaluate  $k_{\text{true}}$  which is the object of the exercise. This however can be achieved by an iterative method [8, 10]. This in turn requires a knowledge of  $n_{\text{app}}$ , the apparent real component of the refractive index which is obtained through a Kramers—Kronig transformation from  $k_{\text{app}}$ . A subtractive form of the Kramers—Kronig transform function is used to accommodate the fact that  $k_{\text{app}}$  is known only over a limited range of the spectrum [10, 11, 13]. This in turn requires a knowledge of the real component of the refractive index ( $n_r$ ) at one wavenumber, obtained by an independent experiment. For this, a small series of refractive index measurements is obtained by attenuated total reflection (a.t.r.) measured on a reflection spectrophotometer designed and constructed for this purpose [14].

The computer programs generated in the course of this work were written in IBM FORTRAN 4 and have been published together with explanatory notes and test data [15, 16]. They are also available on magnetic tape. A statistical error propagation analysis was also done; the initially estimated errors in the cell thickness and transmission measurements and the computational errors associated with the Kramers—Kronig integrations were computed at each data point. In practice the spectra are measured over a range of different cell thicknesses and the computer selects the preferred measurement on the basis of the error propagation algorithm which is incorporated as a sub-routine into the final data reduction program [13].

## RESULTS

A subject of interest has been a survey of materials as possible reference standards for infrared intensity measurements and band shapes. Several common organic liquids suitable for this purpose have been examined. These are listed in Table 1. For various technical reasons, four of these were subsequently discarded but optical constants for the other thirteen have recently been published. These were measured at intervals of  $0.5 \text{ cm}^{-1}$  over the range  $4200\text{--}250 \text{ cm}^{-1}$ . Both the " $n_{\text{true}}$ " and " $k_{\text{true}}$ " curves are published as computer-generated plots, together with tables of the values of the  $k_{\text{true}}$

TABLE 1

Organic liquids selected for a survey study of optical constant measurements in the infrared<sup>a</sup>

1. Cyclo-C <sub>6</sub> H <sub>10</sub>	7. <sup>b</sup> CHBrCl <sub>2</sub>	13. <sup>b</sup> C <sub>6</sub> H <sub>5</sub> F
2. CH <sub>3</sub> -NO <sub>2</sub>	8. <sup>b</sup> CHBr <sub>3</sub>	14. C <sub>6</sub> H <sub>5</sub> Cl
3. CH <sub>3</sub> -CN	9. CBrCl <sub>3</sub>	15. C <sub>6</sub> H <sub>5</sub> Br
4. CH <sub>2</sub> Br <sub>2</sub>	10. CCl <sub>4</sub>	16. C <sub>6</sub> H <sub>5</sub> I
5. CH <sub>2</sub> Cl <sub>2</sub>	11. C <sub>6</sub> H <sub>6</sub>	17. C <sub>6</sub> F <sub>6</sub>
6. <sup>b</sup> CH <sub>2</sub> BrCl	12. C <sub>6</sub> H <sub>5</sub> ·CH <sub>3</sub>	

<sup>a</sup>Range 4200–250 cm<sup>-1</sup> at 0.5-cm<sup>-1</sup> intervals. <sup>b</sup>Rejected for various technical reasons.

maxima and the  $n_{\text{true}}$  extrema with uncertainty estimates based on the error propagation algorithm. In addition, the complete data defining the band contours at 0.5-cm<sup>-1</sup> intervals are available on magnetic tape.

Some similar optical constant data for free standing films of three industrial polymers (Lexan, Trogamid-T, Saran) have also been published [12].

### Conclusion

These results should be of interest even to those with no direct involvement with infrared spectrophotometry, as an example of the general thesis emphasized in the Introduction. This warning of the dangers inherent in using sophisticated measuring instruments without careful analysis of the quantity that is actually being measured becomes particularly important when the data are later subjected to extensive computer processing. These studies were begun in 1967 shortly after acquisition of our first infrared spectrophotometer with facilities for digital data logging on punched paper tape [2]. The work was started rather casually and, in retrospect, rather naively; certainly the ramifications of the problem embarked upon were not anticipated.

### REFERENCES

- 1 R. N. Jones, *J. Jpn. Chem.*, 21 (1967) 609.
- 2 R. N. Jones, *Pure Appl. Chem.*, 18 (1969) 303.
- 3 P. R. Griffiths, C. T. Foskett and R. Cubbello, *Appl. Spectrosc. Rev.*, 6 (1972) 31.
- 4 J. Bratož, J. Rios and Y. Guissani, *J. Chem. Phys.*, 52 (1970) 439.
- 5 J. H. R. Clarke, in R. J. H. Clark and R. E. Hester (Eds.), *Advances in Infrared and Raman Spectroscopy*, Vol. 4, Heyden, London, 1975, Chap. 4.
- 6 H. Tadokoro, *Structure of Crystalline Polymers*, Wiley, New York, 1975, Chap. 5.
- 7 R. P. Young and R. N. Jones, *Chem. Rev.*, 71 (1971) 219.
- 8 R. N. Jones, D. Escobar, J. P. Hawranek, P. Neelakantan and R. P. Young, *J. Mol. Struct.*, 19 (1973) 21.
- 9 J. P. Hawranek, P. Neelakantan, R. P. Young and R. N. Jones, *Spectrochim. Acta*, Part A, 32 (1976) 75.
- 10 J. P. Hawranek, P. Neelakantan, R. P. Young and R. N. Jones, *Spectrochim. Acta*, Part A, 32 (1976) 85.
- 11 J. P. Hawranek and R. N. Jones, *Spectrochim. Acta*, Part A, 32 (1976) 99, 111.

- 12 G. K. Ribbegård and R. N. Jones, *Appl. Spectrosc.*, 34 (1980) 638.
- 13 T. G. Goplen, D. G. Cameron and R. N. Jones, *Appl. Spectrosc.*, 34 (1980) 652, 657.
- 14 D. G. Cameron, D. Escolar, T. G. Goplen, A. Nadeau, R. P. Young and R. N. Jones, *Appl. Spectrosc.*, 34 (1980) 646.
- 15 D. G. Cameron, J. P. Hawranek, P. Neelakantan, R. P. Young and R. N. Jones, *Computer Programs for Infrared Spectrophotometry*, Vol. VI, Programs XLII—XLVII, N.R.C.C. Bulletin No. 16, 1977\*.
- 16 D. G. Cameron, D. Escolar, T. G. Goplen and R. N. Jones, *Computer Programs for Infrared Spectrophotometry*, Vol. VII, Programs XLVIII—L, N.R.C.C. Bulletin No. 17, 1977\*.

---

\* Available at nominal cost from the Publication, Sales and Distribution Division, Building M-58, National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6.

## COMBINATION OF ANALYTICAL SPECTROMETERS AND SPECTROSCOPIC DATA BASES

KOGORO MAEDA\* and YASUJI KOYAMA

*Electrotechnical Laboratory, 1-1-4, Umezono, Sakuramura, Ibaraki-ken (Japan)*

KAZUO SATO

*Japan Information Processing Service Co., Ltd., 1-4, Kabutocho, Nihonbashi, Chuo-ku, Tokyo (Japan)*

SHIN'ICHI SASAKI\*\*

*Miyagi University of Education, Aoba, Aramaki, Sendai (Japan)*

(Received 23rd January 1981)

### SUMMARY

Most search systems receive input data through manual keyboards in remote terminals for searching spectroscopic data bases for candidate compounds corresponding to unknown samples, but manual input systems are slow, subject to human error and mentally fatiguing. This paper records preliminary attempts to transmit spectroscopic data (peak heights and positions) directly to the data base without manual input and to let the central computer retrieve the best-fitting data and transmit them to the remote terminals. First the spectroscopic data were digitized, exact correspondence between the intensities and peak positions being maintained. Chart data were digitized automatically; tests were made for n.m.r. and i.r. data. Direct connexion of an infrared spectrometer to the data base was also tested. Error corrections were made during the transfer processes to ensure exact spectroscopic information. The data were also concentrated for brevity and minimization of error. Two re-transmission systems were tested. In the final step, peak positions and heights were extracted from the transferred data, and the results were used for the search of the i.r. data base. Performance tests showed a considerable degree of success.

Infrared spectroscopy is extremely valuable in the structural analysis of organic substances, e.g., in identification of functional groups. For analytical purposes, the spectra of unknown samples are often compared with those of known samples by the pattern-matching method, visually or by computer techniques with complicated algorithms. In both cases, standard spectra of known samples are necessary, and the importance of spectral data collections has long been recognized.

After Abney and Festing [1] collected the first 48 photographically measured infrared (i.r.) spectra in 1881, several collections of i.r. spectra were published [2-4] and the potentiality of spectral collections for analytical

\*\*Present address: Toyohashi University of Technology, 1-1, Hibariga-oka, Tenpaku-cho, Toyohashi, Japan.

purposes was noted. The importance of having large collections was recognized by the A.S.T.M. E-13 Committee in the early 1950's, and comprehensive indexes gradually developed, culminating in the Wyandotte-A.S.T.M. spectral data file [5], which contains about 145,000 spectra and is widely used although its inadequacies are well known. Some reconstruction is now in progress [6]. Many computer-aided systems for i.r. spectra identification have now been described, almost all of which are based on the Wyandotte-A.S.T.M. File. These programs have mostly been written so as to compensate for the deficiencies of this File.

Although n.m.r. spectra are nearly as important for analytical purposes as i.r. spectra, n.m.r. data bases are less widely available than i.r. data bases. Of course, simultaneous usage of both types of spectra is more effective for analytical purposes than either technique alone.

In every case, spectral information (peak positions and intensities) is obtained from the measured chart data of unknown samples and used for identification of the samples. To date, the spectral information required by computer programs has usually been input manually from keyboards at remote terminals, the peak positions and intensities being input generally through ASCII characters. A mark card system is also available, the data marked manually on the card being input via a mark card reader [7]. Manual input systems, however, are slow, always subject to human error, and mentally fatiguing.

The aim of the work described here is to send the spectroscopic data directly to the data base without manual input of peak heights and positions and to let the central computer retrieve the best-fitting data and transmit it to the remote terminals. In achieving these purposes, there are at least four essential steps. First, the analytical chart data must be digitized, automatically if possible, so that it can be read and processed by computer systems, and it must be confirmed that the digitized data correspond exactly to the chart data and can be reconstructed to give the analog chart data. Further, it must be possible to send the digitized data to special data centers, with the spectral information intact. If this can be done, then chart data already accumulated in laboratories can be utilized. Secondly, it is essential to prove that these digitized data can be used reliably to establish a list of candidate names for unknown compounds. Thirdly, real tests of the system involve sending digitized data to a real data base far distant from the laboratory through telephone lines, to check the retrieval of the best-fitting data. Finally, direct connexion between the spectrometer in the laboratory and the distant data base is needed.

Experiments conducted along these lines are reported in the following paragraphs.

#### DIGITIZATION OF ANALOG CHART DATA [8]

The system used is shown in Fig. 1. A tentative data center was set up at a computer center located about 60 km north of the laboratory in Tokyo.

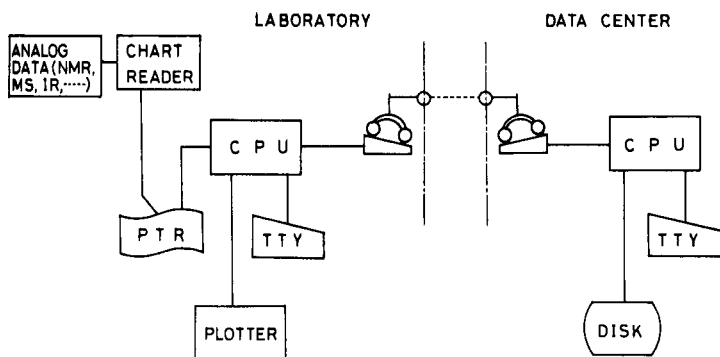


Fig. 1. Digitization, transference and communication system between the laboratory and data center terminals.

Connexions between the two terminals were made through acoustic couplers and the public telephone lines at a rate of 300 bps. The computer system at the laboratory was a NOVA-01 minicomputer with 32K word memory, paper-tape reader, TTY and a plotter; at the data center, another NOVA-01 minicomputer was used with a disk memory and other normal facilities.

As chart data, n.m.r. data were preferred for testing, because the peak profiles are usually more complicated than those of i.r. spectra. The n.m.r. chart paper used had an effective area of 250 mm  $\times$  360 mm for data recording and the analog data were recorded with red or black ink. Recorded data were digitized automatically by using an autochart reader TCR-80 (JEOL); the recorded line was traced with an optical moving sensor by following the change of light transmitted through the chart paper. The optical sensor (0.3-mm diameter light) was moved along the abscissa ( $x_i$  axis) from the top left-hand side to the right-hand side of the paper in 0.1-mm steps and a distance  $y_i$  from a prefixed line (abscissa, base line) was read automatically by the position of the sensor at each step. The microcomputer system processed these position data and output the values of  $y_i$  in 4 ASCII digits. This procedure ensured exact correspondence between  $x_i$  and  $y_i$  and only  $y_i$  values were output on paper tapes and also shown on a small CRT for monitoring. The combination of  $x_i$  and  $y_i$  is easily obtained by counting the distance of the  $x_i$  value from the first value ( $y_1$ ) of  $y_i$  corresponding to  $x_1$ , and combining the counts with the  $y_i$  value. Complete digitization by using only the  $y_i$  values is useful because it allows the transfer time to be halved. The shorter the transfer time, the less trouble is likely during the transfer.

Some  $^1\text{H}$ -n.m.r. chart data are shown in Fig. 2A as an example. The digitized data from the chart on the paper tape were stored tentatively in the memory of the CPU, and transmitted to a plotter for the test. The result obtained by the plotter confirmed that the autochart reader system worked well; it took only 3 min for digitization of Fig. 2A.



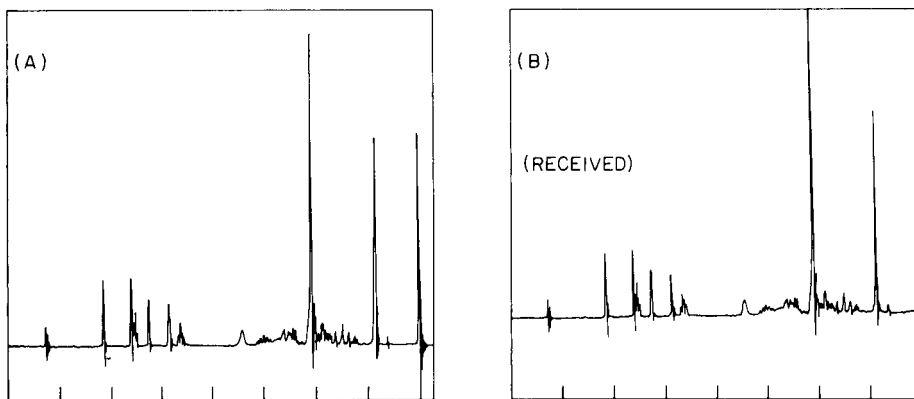


Fig. 2. (A)  $^1\text{H}$ -n.m.r. chart data for the derivative of a natural product, carpesiolide monoacetate (100 MHz). (B) Data received at the laboratory with the protection procedure.

After some communications between the laboratory computer and the data center computer, the stored data were sent to the latter, confirming ready states for sending and receiving between the two terminals. However, public telephone lines are usually subject to noise signals, and procedures for protection from this noise are essential for such transfers of exact chart data. Without a noise-error correction, transmitted data were sometimes completely lost. Initially, data were sent to the center, stored in the disk, and then returned to the laboratory on command from the laboratory computer. In the worst cases, reconstructions of the analog data were hardly possible. The automatic error correction system normally used in large-scale computer systems, and sometimes called ERC, was therefore added to the present system. In the correction system, 5 check bits were used for one-word data of 16 bits, and a one-bit error correction in the one word was possible by using a 5-bit fault code. One word was used as check bits for a 3-word data element. For the case of errors exceeding one bit, block data consisting of 256 words were retransmitted automatically. This two-stage error correction system made it possible to receive correctly at the laboratory data sent earlier to the data center, as shown in Fig. 2B. With noise correction, 4 min was needed for the transmissions.

For more complicated spectra such as Fig. 3A, the correction system was still quite effective. The  $^{13}\text{C}$ -n.m.r. spectrum shown in Fig. 3A is for thio-strepton [9]. The received data, transmitted from the laboratory to the center and back, were plotted as shown in Fig. 3B. Examination proved that the relative heights of the peaks in the transmitted data could be reconstructed within 1% deviation from the originals for both Figs. 2B and 3B.

All the experiments done have confirmed the usefulness of the present system for digitization of spectroscopic chart data and transmission of the data through public telephone lines. With this system, spectroscopic chart data accumulated in distant laboratories can be utilized. Furthermore,

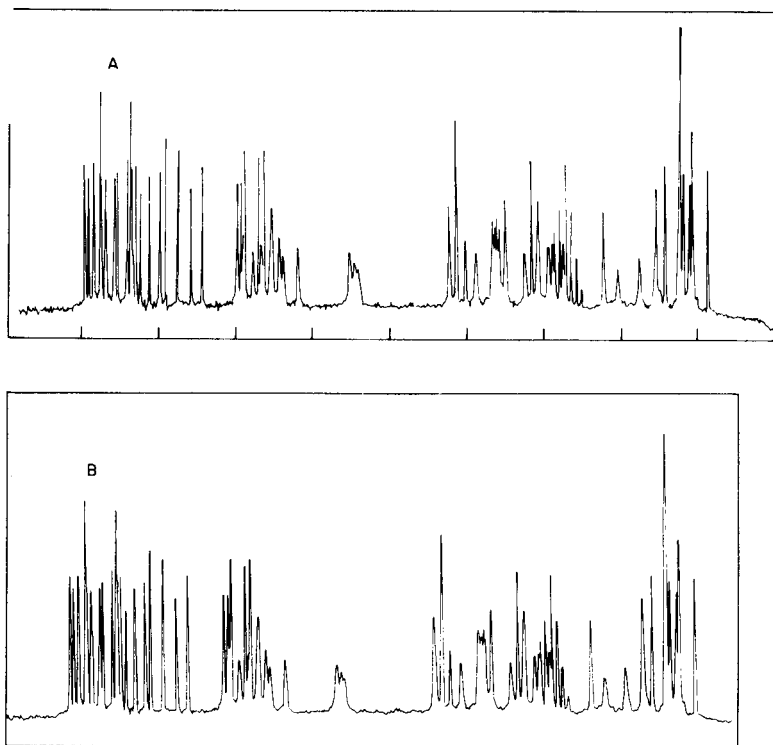


Fig. 3. (A)  $^{13}\text{C}$ -n.m.r. spectrum of thiostrepton,  $\text{C}_{77}\text{H}_{45}\text{N}_{19}\text{O}_{16}\text{S}_5$ . (B) Received spectrum of thiostrepton.

spectroscopic data can be stored in a form directly usable for processing by computer systems.

#### COMBINATION OF DIGITIZED DATA WITH CHEMICS-F [10]

It was then necessary to establish whether or not the digitized data were reliable and effective input data for searching a n.m.r. data base for candidate names for unknown samples. At the time of these experiments, no n.m.r. data base was easily available in Japan, and so the digitization system was combined with CHEMICS-F, which was under development at the Miyagi University of Education and had a small data base of n.m.r. spectra [11].

The system used for this purpose (Fig. 4) was a combination of the FACOM 230-15 computer containing CHEMICS-F programs and a JEC-6 n.m.r. controller connected to a spectrometer through A/D or D/A converters. CHEMICS-F receives only peak positions, peak heights and peak areas of the n.m.r. data, and the data are provided by JEC-6 as paper tapes. It was therefore necessary to extract from the tapes the required numerical  $y_i$  values provided by the digitization procedures described above. Special

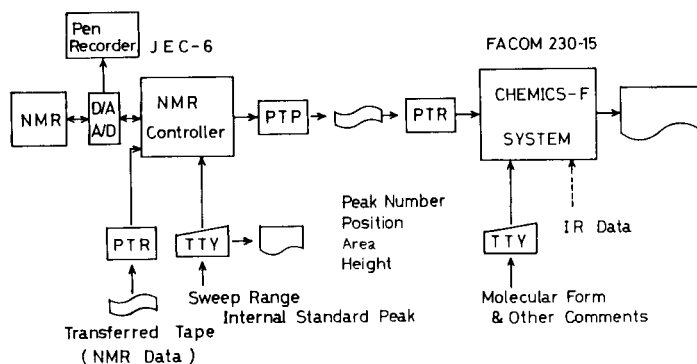


Fig. 4. Combination of the FACOM 230-15 system and the n.m.r. controller system.

programs were set for this extraction, and the extraction procedure was done in the n.m.r. controller JEC-6. By using the digitized n.m.r. data of ethylbenzene and several other compounds, the extraction program was improved by taking the threshold values of the n.m.r. data into account, because in the CHEMICS-F system, threshold treatment is important for obtaining good results. After several tests, the program finally satisfied the requirements of CHEMICS-F. Figure 5 illustrates the system for crotonic aldehyde, a common standard. In this case, digitization was made in the upper edge trace mode, because only the peak positions and the heights are required by CHEMICS-F, and the results were traced by the usual pen recorder, not by plotter. The n.m.r. controller had no communication facilities at the time of these experiments, and the digitized paper tapes were mailed to the processing laboratory where they were read into the n.m.r. controller through a paper tape reader, and subjected to the extraction procedures by the programs refined by the preliminary experiments, which finally gave the peak information shown in Table 1. The original of this table was typed out by a TTY

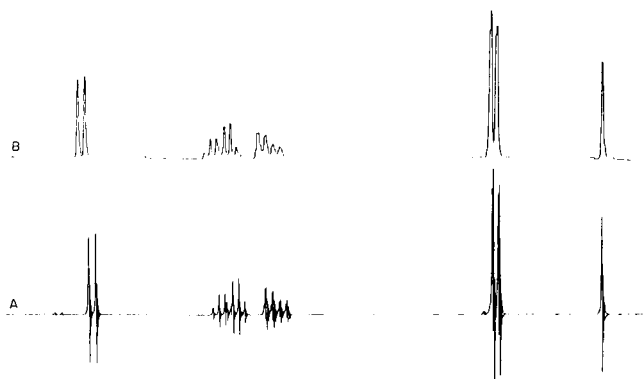


Fig. 5. N.m.r. spectra and the extracted peak profiles for crotonic aldehyde. A, Original spectrum; B, reconstructed spectrum.

TABLE 1

Peak table generated for crotonic aldehyde (cf. Fig. 5)

NO.	POSITION	AREA	HEIGHT	POSITION
1	9153	1086	87	583.9
2	9036	1101	89	575.9
3	6984	206	21	435.9
4	6885	237	21	429.2
5	6756	450	37	420.4
6	6657	451	39	413.6
7	6561	139	13	405.0
8	6201	610	29	382.5
9	6087	501	25	374.7
10	5970	284	15	366.7
11	5856	247	15	359.0
12	2397	2805	159	123.0
13	2298	2518	141	116.2
14	594	1349	105	

connected to the n.m.r. controller, and also punched out as paper tapes to be provided to the CHEMICS-F system. The system composed possible structures of the unknown sample and found candidate names by receiving the peak information on the paper tapes, with the results shown in Table 2. The system indicated that at least two candidates were possible, one of them being crotonic aldehyde.

Another example is shown in Fig. 6 and Table 3. Digitization, mailing, extraction and the illustrations are similar to those for crotonic aldehyde. From the data shown in Table 3, the CHEMICS-F system generated 9 candidates (Table 4) which included the original sample, 3,5,5-trimethylhexanol.

The two examples presented proved that digitized data are certainly capable of being effective input data for generating the candidate names of unknown compounds.

TABLE 2

Components and structures generated by CHEMICS from the peak table (Table 1)

COMPONENT SET NO. 1				COMPONENT SET NO. 2			
1	33	CH <sub>3</sub>	(D)	1	33	CH <sub>3</sub>	(D)
2	118	-CH=<OLEFIN>		2	118	-CH=<OLEFIN>	
3	118	-CH=<OLEFIN>		3	118	-CH=<OLEFIN>	
4	126	-OH	(D)	4	136	-CHO	(D)
5	145	=C=<ALENE>		5	189	<D>	
6	189	<D>		STRUCTURE NO. 2			
7	189	<D>		STRUCTURE NO. 1			
STRUCTURE NO. 1				1-2-5-3-4	CH <sub>3</sub> -CH=CH-CHO		
4-3-7-5-6-2-1	HO-CH=C=CH-CH <sub>3</sub>						

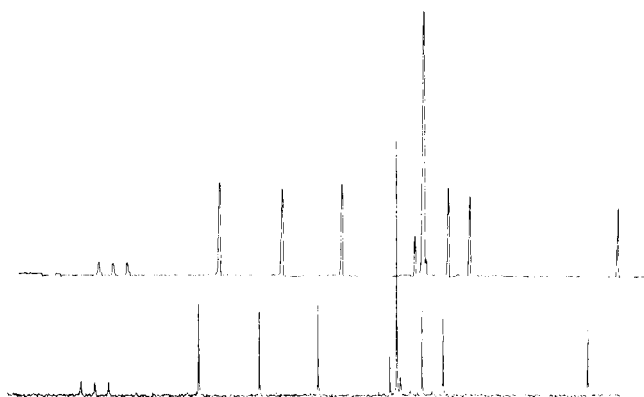


Fig. 6. Recorded and reconstructed n.m.r. spectra for 3,5,5-trimethylhexanol.

TABLE 3

Peak table generated for 3,5,5-trimethylhexanol (cf. Fig. 6)

NO.	POSITION	AREA	HEIGHT	POSITION
1	8949	126	13	
2	8724	124	13	
3	8490	110	13	
4	6966	1012	99	60.9
5	5955	826	91	51.4
6	4980	861	97	42.3
7	3792	290	39	31.1
8	3651	3255	279	29.8
9	3624	121	17	
10	3249	719	91	26.0
11	2895	674	85	22.7
12	474	525	67	

#### COMBINATION OF DIGITIZED DATA WITH A REAL DATA BASE [12]

Based on the results outlined above, combination of the digitization procedure with a real data base was examined in order to assess the practical utility of the system. At the time of these experiments, the only spectroscopic data base available in the TSS mode in Japan was IRSPAN80, this is an i.r. spectra data base based on the Wyandotte-A.S.T.M. File with a retrieval system developed by the Japan Information Processing Service.

Figure 7 shows the combined system used. The laboratory computer system was almost the same as indicated for Fig. 1, except for the addition of a CRT and a two-drive floppy disk system; at the data center a large

TABLE 4

Simulated output of structures generated by CHEMICS from the peak table (Table 3)

1.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{OH}$	6.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\overset{\text{CH}_3}{\underset{\text{CH}_3}{\text{C}}}-\text{CH}_2-\text{OH}$
2.	$(\text{CH}_3)_3\text{C}-\overset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{OH}$	7.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{O}-\text{CH}_3$
3.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\overset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{CH}_2-\text{OH}$	8.	$(\text{CH}_3)_3\text{C}-\overset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{CH}_2-\text{O}-\text{CH}_3$
4.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\text{CH}_2-\overset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{OH}$	9.	$(\text{CH}_3)_3\text{C}-\text{CH}_2-\overset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{O}-\text{CH}_3$
5.	$(\text{CH}_3)_3\text{C}-\overset{\text{CH}_3}{\underset{\text{CH}_3}{\text{C}}}-\text{CH}_2-\text{CH}_2-\text{OH}$		

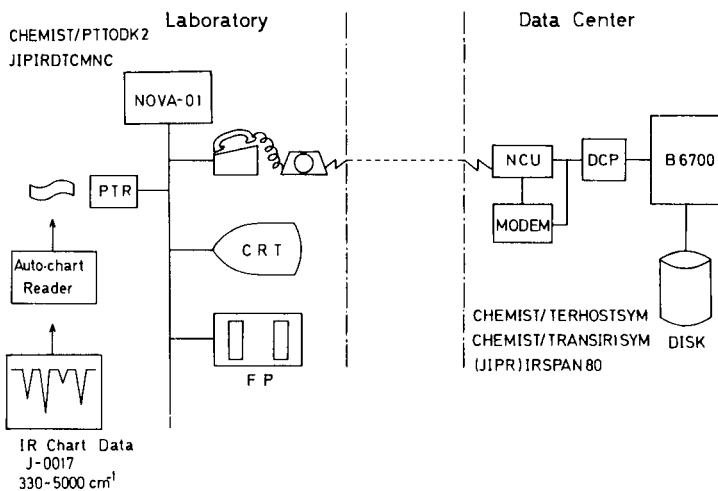


Fig. 7. Combination of the laboratory system with the IRSPAN80 data base.

Burroughs B6700 computer was used with large disk system and other facilities for TSS operations.

The analog i.r. charts used for the test were all calibrated in wavelength (2.0–30.3  $\mu\text{m}$ ) and typical in size (160 mm  $\times$  285 mm). The recorded i.r. data were digitized on paper tapes as described for n.m.r. spectra. The tapes were directly read into the disk memory through the paper-tape reader and a MT system at the data center (not through telephone lines) for the first tests of the suitability of the digitized data for generating candidate names

for unknown samples from a real data base. The ASCII data read into the system were converted to EBCDIC data and then subjected to the extraction procedure for generating candidate compounds from the IRSPAN80 Data Base. The programs were \$CHEMIST/TRANSIR for the extraction procedure and (JIPR) IRSPAN80 for the generation.

For testing, the spectrum of acetophenone, a typical i.r. example, was used. Through the series of the procedures described above, eleven compound names were finally generated from the data base, ten of which corresponded to acetophenone. In searching, the wavenumber latitude was set at  $20\text{ cm}^{-1}$ . The eleven names were decreased to eight when the latitude was set at  $10\text{ cm}^{-1}$ , and the eight candidates were all similar to acetophenone.

These preliminary experiments showed that the digitized data on paper tapes were capable of providing effective input data for the retrieval system of the real data base.

The digitized data on paper tape were then used for further experiments in which laboratory data were transferred to the data center through telephone lines. The digitized data on the tapes were read into the CPU system by the paper-tape reader and stored in the floppy disk system at the laboratory. These data were then converted in a more reduced form by changing the 4 ASCII digits to 2-byte data and by condensing the spectroscopically redundant parts in the original data. This was done by the PTTODK2 program in the laboratory computer and allowed the transfer time to be approximately halved, e.g., from 9 min to 3–4 min in the case of acetophenone. The reduced data were stored in floppy disks with different file names, and at the transfer time the data were moved into the main memory of the CPU system. The transfer operation was done with the JIPIRDTCMNC program, which communicated with the data center and also protected from noise during transfer. The protection procedure involved a bit check system of the usual type: the results of the accumulation of binary bits of ASCII characters were compared before and after transfer for every record consisting of 80 ASCII characters representing the spectroscopic data in condensed forms; the retransmission of records from the data center was also monitored. All these operations were done with the floppy disk system in the NOVA-01 CPU system, FDOS.

The transferred data were processed by the CHEMIST/TRANSIR1SYM program, being reconstructed into the original chart form, and then smoothed, differentiated (first and second) to obtain exact spectroscopic peak positions and heights. The results were illustrated by the program at the center, if required. Figure 8 shows the example of acetophenone; comparison with the source data showed that the transfer process worked almost perfectly.

The program (JIPR)IRSPAN80 generated the candidate compound names by using the extracted data on peak positions and heights, and finally returned the answers to the laboratory. In this case, the wavenumber latitude was set at  $10\text{ cm}^{-1}$  and 8 candidates were generated, all of which were similar

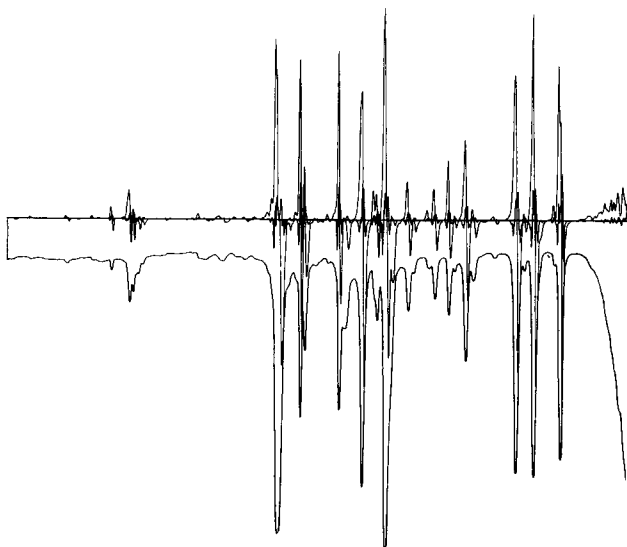


Fig. 8. Transferred data for acetophenone smoothed and differentiated at the data center. From bottom to top: original curve, smoothed curve, first differential, second differential.

to acetophenone as in the preliminary experiment. The connection time was about only 5 min for all the procedures. The results obtained indicated that the transfer system worked well.

In a further example, the i.r. spectrum of polystyrene film was used. The chart was processed as described above for acetophenone, and finally subjected to the search procedure by (JIPR)IRSPAN80. The reconstructed data are illustrated in Fig. 9 together with the first and second differentiations. The final results gave 19 candidates, 8 of which were similar to polystyrene. When the search procedure was done manually with the original data, 33 candidates were produced, only 5 of which corresponded to polystyrene. When the latitude was widened to  $20\text{ cm}^{-1}$  from  $10\text{ cm}^{-1}$ , the (JIPR)-IRSPAN80 gives 29 candidate names, only eight of which were correct. Manual input gave 52 names, only six of which were right.

These results indicate that the present digitization system can give more accurate and reliable results than manual input. The system avoids human errors in observing peak profiles from the recorded spectra and eliminates mental pressure and fatigue.

Some further experiments with samples of different concentrations in different solvents showed that appropriate concentrations in suitable solvents were necessary, as in manual techniques, for obtaining good final results through the digitization and searching procedures.

It can be concluded from all the above experiments that the digitization system may be used fairly reliably in combination with a real data base. With this system, it is possible to generate candidate compound names for unknown samples from remote data bases containing appropriate i.r. or n.m.r. data.



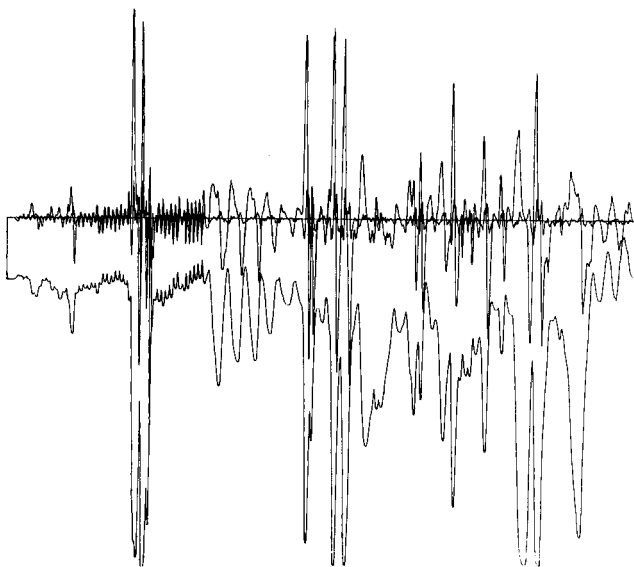


Fig. 9. Transferred, smoothed, and differentiated data for polystyrene film.

#### ON-LINE COUPLING OF AN INFRARED SPECTROMETER TO A REAL DATA BASE [13]

The above sections have shown that candidate compound names for unknown samples can be retrieved from a large data base simply by sending the spectroscopic chart data through public telephone communication lines. Although the processes used were off-line, spectroscopic chart data accumulated in different laboratories could be used.

For new data, however, real-time combination, i.e., on-line coupling, should be examined. Direct combination of a spectrometer with a data center through telephone lines was therefore examined. For this purpose, the on-line system shown in Fig. 10 was used. A Perkin-Elmer i.r. spectrometer Model 283 was employed with a Micro-NOVA microcomputer and two-drive floppy disk systems. The TTY, the acoustic coupler and the i.r. spectrometer were all combined through the RS232C mode, and the i.r. data from the spectrometer were directly transferred to the microcomputer system, condensed and further stored in the floppy disks. The stored data were then transferred to the data center, if required immediately, i.e., on-line. Micro-NOVA had a 24K-words memory in the present system, and all the computer operations controlled by the disk operating system, which monitored and controlled all the I/O procedures over all the peripheral devices. After finishing one operational command, the system control always returned to the disk operating system. This is very convenient and effective for repeating i.r. scanning and for generating candidate compounds.

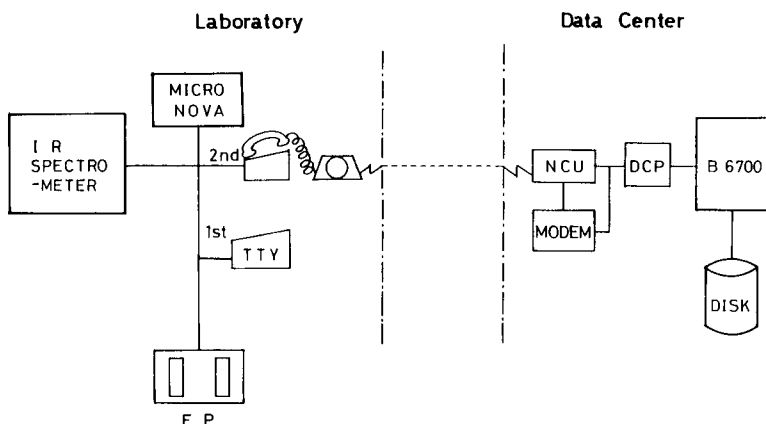


Fig. 10. System for direct on-line coupling between the laboratory and the data center.

The Perkin-Elmer 283 spectrometer is microcomputer-controlled; scanning conditions presented by the ordinate and abscissa switches are read by the microcomputer. Scanning then starts from  $4000\text{ cm}^{-1}$  with a scanning interval of  $1\text{ cm}^{-1}$ . The intensity data, represented in 4 ASCII digits and grouped every ten data points in descending wavenumbers down to  $200\text{ cm}^{-1}$ , were provided by the present microcomputer system. In typical cases, 3800 data points were taken from  $4000\text{ cm}^{-1}$  to  $200\text{ cm}^{-1}$ , and these were converted to about 20 kbyte ASCII characters by the microcomputer installed in the spectrometer. The ASCII data were then transferred to the Micro-NOVA system at 1200 bps (maximum in this spectrometer system), condensed to about a quarter, and transferred to the data center or to storage in a floppy disk. Procedures for generating candidate names for the sample by using condensed data were almost the same as in the off-line tests, except that all the laboratory-terminal procedures with the Micro-NOVA system were controlled by the disk-operating system.

After communications had been suitably established between the laboratory terminal, tentatively located in Osaka for the test, and the same i.r. data center in Tokyo as used previously, the system retrieved 6 candidates for the sample of polystyrene film. Reference to the literature showed that three candidates corresponded to polystyrene, and three were sec-butylbenzene, 1,2-diphenylethane and 1,2-diphenylpropane; the wavenumber width for searching was  $10\text{ cm}^{-1}$ . When the width was expanded to  $20\text{ cm}^{-1}$ , 42 candidates were retrieved, 11 of which corresponded to polystyrene. The time necessary to obtain the final answers was about 5 min from the start of communication with the data center, almost the same as in the previous case.

All the results obtained have certainly shown that the on-line coupling system is effective in generating candidate names of unknown samples by sending the i.r. data directly from the i.r. spectrometer to the data center.

### Conclusions

At the present stage of development, it is possible to utilize i.r. and n.m.r. spectroscopic data, either as chart data forms or in a direct supply mode from the spectrometers, for identification of unknowns through public telephone communication lines even at places remote from data centers. A new route is now open, by which spectroscopic data can be exchanged through public telephone lines between research workers and research organizations. By such exchange and communication, common spectroscopic data bases will become available in computer networks for storage and retrieval.

The authors are greatly indebted to Mr. T. Sakuma, Watanabe Measuring Instrument Co., and Mr. Y. Masuda, JEOL Co., who helped to digitize chart data; Mr. Y. Tezuka and Mr. H. Onishi, Nippon Data General Co., for the transfer of digitized data; Prof. H. Abe who provided the program for combining the digitized data with CHEMICS-F, and the acetophenone and polystyrene data; and Mr. T. Nomura, Abe Trading Co., for help in combining the Perkin-Elmer Model 283 with IRSPAN80.

### REFERENCES

- 1 W. de W. Abney and R. E. Festing, *Phil. Trans. R. Soc. London*, 172 (1881) 1530.
- 2 W. W. Coblenz, *Carnegie Institute of Washington, Publication No. 35* (1905).
- 3 J. Lecomte, in V. Grignard (Ed.), *Traité de Chimie Organique*, Vol. 2, Masson et Cile, Paris, 1936.
- 4 R. B. Barnes, R. C. Gore, U. Liddell and V. Z. Williams, *Ind. Eng. Chem., Anal. Ed.*, 15 (1943) 659.
- 5 ASTM, AMD-31, 32, 34-S15 (1974).
- 6 C. D. Craver, E. M. Kirby and R. Norman Jones, *Proc. 7th Int. CODATA Conf. Kyoto*, 1980.
- 7 Y. Osada, T. Hashimoto, S. Kogure and T. Konishi, *Eng. Chem.*, 23 (1978) 456.
- 8 T. Sakuma, Y. Koyama, Y. Masuda, Y. Tezuka, K. Maeda and S. Sasaki, *Proc. Ann. Meeting Chem. Soc. Jpn.*, 1976.
- 9 K. Tori, K. Tokura, K. Okabe, M. Ebata, H. Otsuka and G. Lukacs, *Tetrahedron Lett.*, (1976) 185.
- 10 H. Abe, Y. Koyama, K. Maeda and S. Sasaki, *Proc. Ann. Meeting Chem. Soc. Jpn.*, 1977.
- 11 S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito and T. Yamasaki, *J. Chem. Inf. Comput. Sci.*, 18 (1978) 211.
- 12 Y. Koyama, K. Sato, H. Abe, K. Maeda and S. Sasaki, *Proc. Ann. Meeting Chem. Soc. Jpn.*, 1978.
- 13 T. Nomura, K. Sato, Y. Koayama, K. Maeda and S. Sasaki, *Abstracts, ACS/CSJ Chemical Congress, Honolulu, Hawaii, 1979, CHIF 45.*

## AN ARTIFICIAL INTELLIGENCE SYSTEM FOR COMPUTER-AIDED MASS SPECTRA INTERPRETATION OF SATURATED ALIPHATIC MONOHYDRIC ALCOHOLS

CHU DAMO, CHANG DACHUN and KUAN TESHU

*Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian  
(People's Republic of China)*

CHEN SHAOYU

*Dalian Shipyard, Dalian (People's Republic of China)*

(Received 23rd January 1981)

### SUMMARY

The mass spectra of saturated monohydric alcohols are used in structure elucidation by means of computer programs of the artificial intelligence type. The structure generator is constrained by a structure predictor to show the sub-group structure in a fairly definite area, and spectral characteristics are also predicted. According to the commands of the structure predictor, sub-groups conforming to the mass spectra studied are then combined into molecular structures which are ranked for output. The system was tested for 218 mass spectra of 115 alcohols from different laboratories: all the correct answers were included in the 6 highest scores, with 78% for the highest score and 91% for the highest 2 scores.

The mass spectral library matching technique, has been widely used in structure identification [1]. Its chief drawbacks are that it does not use the abundant empirical rules of mass spectrometry, and that its elucidation ability is closely related to the presence of the spectrum in the data base. These drawbacks lead to insufficient use of correlations between mass spectra and structures, i.e., much of the capacity of mass spectra in elucidating organic molecular structure is neglected.

Depending on the fragmentation of ionic molecules and the influence of substituents on mass number or intensity, chemists interpret the mass spectra of complex molecules by referring to those of related compounds. Interpretation may sometimes need other structural information such as that available from i.r. or n.m.r. spectroscopy.

In studying the correlation of molecular structures and their mass spectra, the application of mass spectra fingerprints for identification of sub-group structure is fairly effective owing to its structural sensitivity. The elucidation process will be described in detail elsewhere [2, 3]. Here, only the interpretation of mass spectra of saturated aliphatic monohydric alcohols and some results are outlined as an example.

The advantages of the new process are as follows. First, it is suitable for structure elucidation of large molecules containing up to 30 carbon atoms. Second, the only input needed is low-resolution mass spectral data. Third, other structural information, e.g., i.r. or n.m.r. is not usually required. Fourth, the unknown is always correctly defined by one of a few ( $>6$ ) candidates. Finally, mass spectra measured with instruments of different models can be correctly interpreted. This flexibility makes the program applicable in any laboratory.

## METHOD

In the program used here for mass spectral classification, an unknown chain compound may be identified as one of 318 kinds of chemical structures [3]. The mass spectra are then used for structure elucidation.

The interpretation program for saturated aliphatic monohydric alcohols comprises programs for molecular weight determination, for structure prediction and for structure generation. Every sub-program is worked out according to the rules of mass spectrometry. Briefly, the overall operation is as follows. First, the molecular weight is determined. Second, the monohydric alcohols are classified by the structure predictor program into the following fundamental structural models: primary ( $R'CH_2OH$ ), secondary ( $CH_3R'CHOH$ ,  $R'R''CHOH$ ,  $R'R'CHOH$ ) and tertiary ( $R'R''R'''COH$ ,  $(R')_2R''COH$ ,  $R'(R'')_2COH$ ,  $(R')_3COH$ ), where  $R'$ ,  $R''$  and  $R'''$  represent alkyl radicals. They may have equal mass, but may be in different skeletons. If the radicals are of different mass, then  $R' > R'' > R'''$ . Third, the corresponding sub-group structures are output by the structure generator in the light of the commands of a structure predictor and possible sub-groups are combined to give a reasonable structure in agreement with the unknown spectrum. Reasonable structures are stored. If no reasonable candidate structure is presented, then a feedback loop to the molecular weight determination comes into action and the predictor and generator programs are repeated. These feedback loops provide other ways of interpretation in order to obtain reasonable structures when there has been some mistake of interpretation or when no candidate structure is output. The process of interpretation will be repeated until there is no more replenishment. Finally, all the results are accumulated, and a list of candidate structures is output with their order of reliability.

All computer programs, in ALGOL, were run on the DJS-6 model computer (cycle time 20  $\mu s$ ). The whole program needs 35K words. Interpretation of 218 mass spectra takes about 1 h.

The low-resolution mass spectral data used were taken from MSDC [4] and from the Atlas of Mass Spectral Data [5]. The total data set used consisted of 218 spectra from 115 compounds.

## RESULTS AND DISCUSSION

In view of the probable variation in mass spectra for the same compound from different instruments, any suitable program must be sufficiently

TABLE 1

The distribution of the results of interpretation

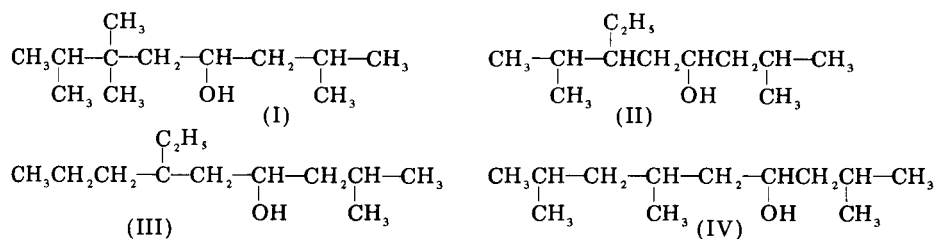
Distribution of candidates			Distribution of correct answers		
No. of candidates	No. of spectra	%	Rank of correct answers	No. of spectra	%
0	1	0.5	0	1	0.5
1	118	55.9	1	165	78.2
2	35	16.6	2	28	13.2
3	16	7.6	3	9	4.3
4	21	9.6	4	6	2.8
5	0	0	5	0	0
6	16	7.6	6	2	1.0
>6	4	1.9			

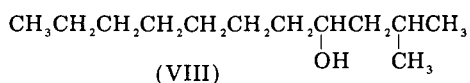
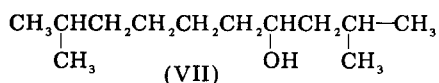
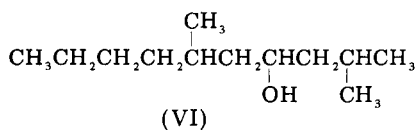
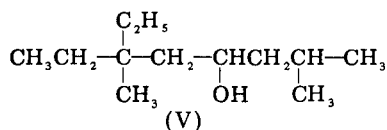
flexible to provide correct interpretation of spectra from various mass spectrometers. The data base used comprised 218 data sheets from different laboratories and included those of the alcohols with the most complex skeletons in the collections [4, 5]. The results of the interpretation are summarized in Table 1. For eight spectra, the program failed to give the answer denoted in the data base. Of these eight failures, seven were due to errors in the spectra, including interference from impurities; these spectra were then eliminated from the data base. The eighth case was a failure of the program.

The results show that the program normally gave the correct answers within a few candidates. Among the 211 spectra used, correct single candidates make up 56% of the total. Correct results make up 78% for the highest score case and 91% for the two highest scores. All the correct results, except one, are included in the first six scores; the exception was one that failed to output the molecular weight of the unknown.

Table 2 gives a comparison of the results obtained for alcohols by the DENDRAL program [6] and the present program. Obviously, the DENDRAL program is in difficulty when used for the elucidation of large molecules, especially those with a complex skeleton. With the present program, correct results can be obtained with much higher probability whether the molecule is simple or complex (Table 2).

With the new program, the spectra of 2,6,8-trimethylnonanol-4 gives the ranked candidates:





The fourth is the correct one. It is interesting to note that in each candidate structure, the isobutyl group always joins the  $\alpha$ -carbon atom and that all the  $\beta$ -carbons are present as methylene. This shows that the structure generator program has a high level of structure specification.

TABLE 2

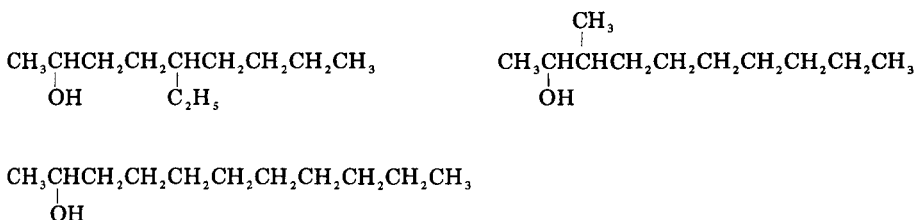
Results of the computer interpretation of some saturated monohydric alcohols

Alcohol	No. of isomers	Number of inferred isomers		
		Dendral program		Present program
		A <sup>a</sup>	B <sup>b</sup>	
n-Butyl	7	2	1	1
sec-Butyl	7	3	2	1
2-Methyl-2-butyl	14	1	1	3
n-Pentyl	14	4	1	1
3-Pentyl	14	1	1	2
2-Pentyl	14	2	1	2
3-Hexyl	32	2	1	2
3-Methyl-1-pentyl	32	8	4	1
n-Hexyl	32	8	1	1
3-Heptyl	72	4	1	2
3-Ethyl-3-pentyl	72	1	1	2
2,4-Dimethyl-3-pentyl	72	3	1	4
3-Methyl-1-hexyl	72	17	6	1
n-Octyl	171	39	1	1
3-Octyl	171	8	1	2
2,3,4-Trimethyl-3-pentyl	171	3	1	6
n-Nonyl	405	89	1	1
2-Nonyl	405	39	1	1
6-Ethyl-3-octyl	989	39	9	6
3,7-Dimethyl-1-octyl	989	211	41	1
2-Butyl-1-octyl	6045	1238	25	1
n-Dodecyl	6045	1238	1	1
3-Tetradecyl	38322	1238	1	1
n-Tetradecyl	38322	7639	1	1
n-Hexadecyl	151375	48865	1	1

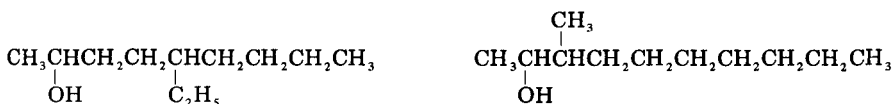
<sup>a</sup>Isomers inferred when only mass spectrometry is used. <sup>b</sup>Isomers inferred when the number of methyl radicals is known from n.m.r. data.

For the interpretation of mass spectra taken from different laboratories, the present program is shown to be highly efficient. Figure 1(a) is taken from AMSD 1160-3 (COC-4791) and Fig. 1(b) from AMSD 1161-3 (DOW-2971). The similarity is only 49% by Euclidean techniques and Biemann's method for spectrum compression [7]. The present program prints the correct skeleton in both candidate lists, although there are many different points between the two spectra.

For AMSD 1160-3, the program gives three candidates:



whereas for AMSD 1161-3, two candidates are printed out:



The correct answer is 5-ethyl-nonanol-2 which has the highest score in both lists.

For the elucidation of structure, it is more reasonable to utilize the correlation of molecular structure and its mass spectrum than simply to identify it

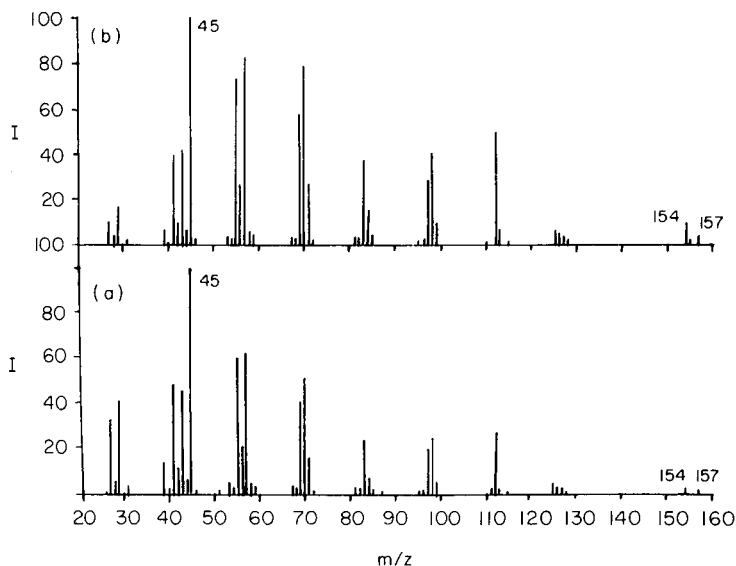


Fig. 1. Mass spectra from different laboratories: (a) AMSD 1160-3; (b) AMSD 1161-3.



with a spectral fingerprint. The differentiation of AMSD 803-3 (DOT-0067) and 801-1 (DOW-1981) is another example of the ability of the present program. Both spectra are designated originally as 3,5,5-trimethylhexanol-1. Their similarity is more than 86%. The program shows that the spectrum in Fig. 2(a) is not 3,5,5-trimethylhexanol-1, but 2,5,5-trimethylhexanol-1 (the only candidate printed out). Obviously, this is correct, because molecular fragmentation of the latter gives  $m/z$  113, while that of the former gives  $m/z$  99.

The program suggested that the spectrum AMSD 993-4 is not in accord with its printed name, 2,7-dimethyloctanol, but should be 3,7-dimethyloctanol. Similarly, with the spectra of AMSD 287-4, 167-5 and 168-5, the printed names should be 2-methylpentanol-2, 2-3-dimethylbutanol-1 and 2-methylbutanol-1, instead of the original names 2-methylbutanol-2, 3-methylbutanol-1 and 3-methylbutanol-1, respectively. There are some other mistakes in the spectra of alcohols in the MSDC and AMSD collections and these spectra were not used as data sheets.

The only failure in interpretation is that of the spectra shown in Fig. 2(b). In this case, no candidate was printed out because the molecular weight could not be derived from the spectrum. It is well known that it is difficult to obtain the molecular ion from many alcohols. The present program has the ability to determine molecular weight, so that low-resolution mass spectra can often be correctly interpreted without molecular weight input. For some primary alcohols with high molecular weights and complex structures, the program may have difficulty in providing correct molecular weights. This problem, which is common to artificial intelligence programs, remains to be solved.

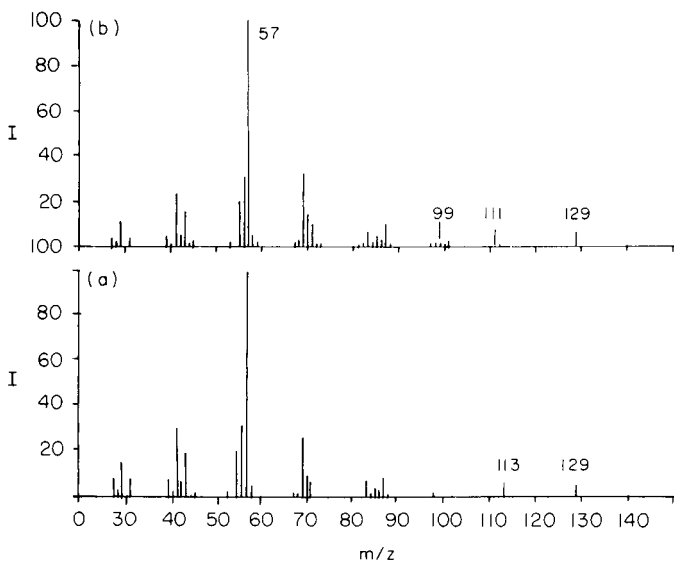


Fig. 2. Mass spectra from different laboratories: (a) AMSD 803-3; (b) AMSD 801-1.

## REFERENCES

- 1 F. W. McLafferty and R. Venkataraghavan, *J. Chromatogr. Sci.*, 17 (1979) 24.
- 2 Chu Damo, Application of sub-group structure on the interpretation of mass spectra. (Unpublished.)
- 3 Chu Damo, Computer-aided mass spectral interpretation, Mass spectral classification of organic chain compounds, Rep. Dalian Inst. Chem. Phys. Chinese Acad. Sci., to be printed (Sept. 1981)
- 4 Mass spectrometry Data Centre: mass spectral interpretation, Mass spectral classification
- 5 E. Stenhagen, S. Abrahamsson and F. W. McLafferty, *Atlas of Mass Spectral Data*, Wiley-Interscience, New York, 1969.
- 6 B. G. Buchanan, A. M. Duffield and A. V. Robertson, in C. W. A. Milne (Ed.), *Mass Spectrometry: Techniques and Applications*, Wiley-Interscience, New York, 1971, p. 121.
- 7 G. T. Rasmussen and T. L. Isenhour, *J. Chem. Inf. Comput. Sci.*, 19 (1979) 179.

## A COMBINED LINEAR AND NONLINEAR FACTOR ANALYSIS PROGRAM PACKAGE FOR CHEMICAL DATA EVALUATION

CLEMENS JOCHUM\* and B. R. KOWALSKI

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

(Received 23rd January 1981)

### SUMMARY

An underlying variable factor analysis program (UVFA) is described. The theories of principal component analysis and nonlinear least-squares projection techniques are outlined and compared. Several applications from various chemical fields are presented which show that a complete analysis of the underlying structure and dimensionality of a chemical data set should always include these nonlinear projection techniques.

Multivariate statistics, originally developed for applications in social sciences, have been more and more applied to chemical data evaluation. In fact, the statistical treatment of chemical data has become a new branch of analytical chemistry, called chemometrics [1].

One of the most powerful methods in chemometrics which has been applied as a "stand-alone" method as well as in combination with other methods is principal component factor analysis. Because communality estimate and iteration are of minor importance for chemical applications, principal component and linear factor analysis are not distinguished here; more detailed information is readily available [2]. Applications range from data reduction problems, interpretation of the underlying structure of a data set to a preliminary treatment of the data bases for a path modelling analysis [3]. Principal component analysis has been applied to, e.g., mass spectral and environmental data, n.m.r. and chromatography data [4].

Principal component analysis assumes a linear relation among the variables. In nature, however, most relations between physical parameters or variables are nonlinear. To overcome this setback of linear factor analysis, algorithms such as nonlinear least-squares multidimensional scaling [5] and parametric mapping [6] for the analysis of the underlying structure of a data base have been developed. So far, no applications of these nonlinear methods to chemical data analysis have been published.

The theory of the different linear and nonlinear methods is explained below, and an interactive program package is described which includes not only principal component factor analysis and rotational methods, but

also nonlinear least-squares projection techniques such as multidimensional scaling, nonlinear and parametric mapping and graphical output routines. The algorithms and the program are demonstrated on two chemical data sets.

## THEORY

The underlying relation of  $n$  variables (e.g., physical parameters like melting point, dipole moment, etc.) of a data matrix,  $Z = (Z_{ij})$   $i = 1, \dots, m$ ;  $j = 1, \dots, n$ , consisting of  $m$  measurements for each variable is to be analyzed. To give the variables equal weight, they are usually scaled to unit variance and zero mean. In a three-variable data set ( $n = 3$ ) the measurement vectors can be represented geographically in a three-dimensional space (Fig. 1).

Factor analysis determines the dimensionality of the hyperspace necessary to represent the data. The first factor  $\lambda_1$  is represented by the longest axis of the hyperspace containing the data, i.e., it represents the largest amount of variance in one dimension. The second vector  $\lambda_2$  is represented by the second longest axis orthogonal to the first one and so on. To obtain the  $r$  factors necessary to represent most of the total variance of the data set, the data matrix  $Z$  is decomposed in a factor weight matrix (factor loading matrix)  $A = (a_{ij})$   $i = 1, \dots, n$ ;  $j = 1, \dots, r$ , and a factor score matrix  $P = (p_{ij})$   $i = 1, \dots, m$ ;  $j = 1, \dots, r$ ; ( $r \leq n$ ):  $Z = P \cdot A^T$ .

The columns of  $A$  are determined by calculating the eigenvectors of the data covariance matrix  $C$  where  $C = Z^T Z$ . The entries  $a_{ij}$  of the factor-loading matrix  $A$  can be considered as the multiple correlation coefficient of the variable  $i$  with the factor  $j$ .

The factor score matrix represents the data in terms of factor coordinates and is calculated from  $P = ZA$ . This transformation is known as the Karhunen-Loeve expansion [7].

As mentioned above, the relation between physical parameter variables is not always linear. It is, for example, possible that the data lie along a curved line or surface (Fig. 2).

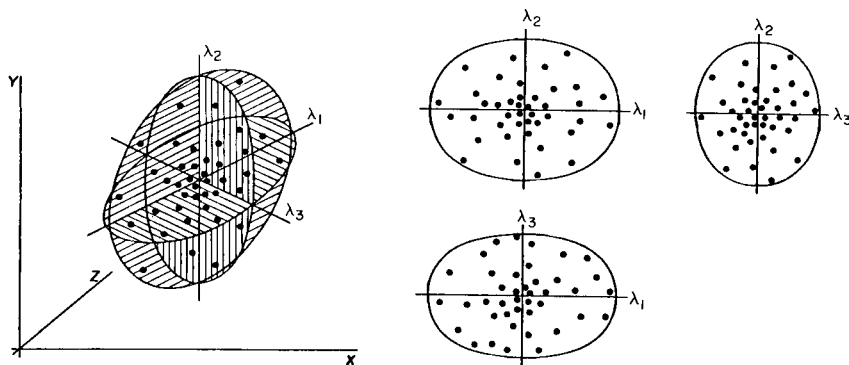


Fig. 1. Three-dimensional representation of measurement vectors.

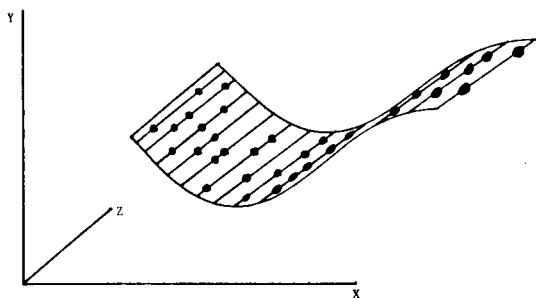


Fig. 2. Three-dimensional data points lying on a nonlinear surface.

Linear principal component factor analysis would still come up with three factors because the variance for all possible three-dimensional orthogonal coordinate systems is greater than zero in any coordinate direction. Yet there are obviously only two underlying nonlinear independent variables. To solve this problem, nonlinear least-squares projection methods have been developed [5, 6].

The distances ( $d_{ij}$ )  $i, j = 1, \dots, m$  between all  $m$  data vectors of  $Z$  are calculated. The data points are then arranged in an  $r$ -dimensional space ( $r < n$ ) such that for the stress  $S$

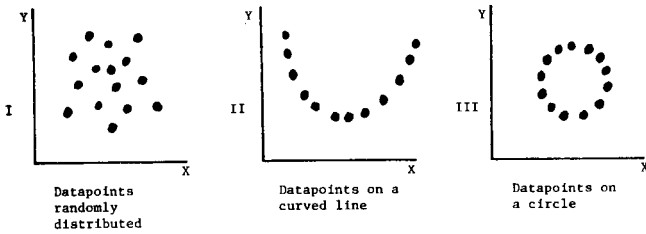
$$S = \left\{ \left[ \sum_{i,j} (d_{ij} - \hat{d}_{ij})^2 \right] / \left[ \sum_{i,j} d_{ij}^2 \right] \right\}^{1/2}$$

is minimal [5]. The  $\hat{d}_{ij}$  terms denote the recalculated distances of the data points in the lower  $r$ -dimensional space. The stress  $S$  thus represents a measure for the goodness of fit of the data vectors projected in the  $r$ -dimensional space compared with their configuration in the original  $n$ -dimensional space.

The different projection techniques differ mainly by a different measure for the goodness of fit. To demonstrate the different applications for principal component analysis, multidimensional scaling [5] and parametric mapping [6], their optimum theoretical results on three different two-dimensional data sets (I, II, III) are shown (Fig. 3).

The parametric mapping algorithm is able to determine a ring-shaped one-dimensional structure of the data because it considers only local environments of data points. Since this method does not look at the global fit of all data points, however, it sometimes ends up with a too small dimensionality.

Although there exist programs for these nonlinear least-squares methods, they are not set up for chemical data bases, they are not input-compatible with each other and they work only as batch programs. Because these programs only include either multidimensional scaling or parametric mapping and no linear factor analysis program, there was a definite need for a combined package. Such a combined underlying variable factor analysis program is described below.



Dataset	Dimensionality found by:		
	Principal component analysis	Multidimensional scaling	Parametric mapping
I	2	2	2
II	2	1	1
III	2	2	1

Fig. 3. Comparison of principal component, multidimensional scaling and parametric mapping analysis.

#### THE PROGRAM UVFA

The underlying variable factor analysis program UVFA consists of a driver routine, a set of utility routines and 21 major subroutines which perform the actual data analysis. Because the data are stored on disk files, only the driver routine and the utility routines have to stay in core during the whole run. The 21 major subroutines can be loaded one at a time. Thus the program usually needs less than 60<sub>8</sub>K words of core to run although it consists of more than 10,000 statements.

The input, output and internal binary files are fully compatible with the pattern recognition program ARTHUR [8].

UVFA can be run interactively or in batch mode and has graphical output routines for Tektronix 4010/4014 terminals, Calcomp plotter or line printer. Figure 4 shows the general setup of the program. PRICO does a principal component analysis with or without communality iteration [2]. MULSCA and PARAMA are the nonlinear least-squares projection routines for multidimensional scaling and parametric mapping. The underlying linear and nonlinear factors can be plotted with the routines PRIPLO (line printer plot), CALPLO (Calcomp plot) or TEKPLO (Tektronix graphics terminal plot). For additional error analysis the linear factors can be back-transformed by calling KATRAN (Karhunen—Loeve Transformation) and BACKTR. The program also assists with the interpretation of the factors by calling ANALYS (ordering of the factors and loadings and performing various tests for finding the intrinsic dimensionality), HIER (performing a hierarchal cluster analysis), ROTOR and ROTSUB for performing various kinds of rotations.

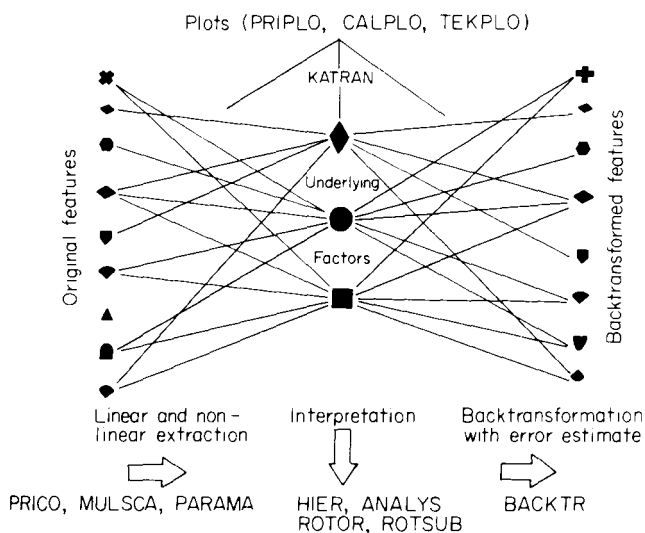


Fig. 4. Schematic diagram of the UVFA program.

There are versions for 60-bit CDC computers and 32-bit DEC VAX computers. The VAX version should be well compatible with other DEC and IBM computers. The whole program is written in FORTRAN, and is available from Infometrix, (P.O. Box 25808, Seattle, WA 98125).

#### APPLICATIONS

Among the various applications, three are discussed in more detail: a mass spectral data set, a constitutional similarity set of chemical compounds, and a data set of physical parameters of biologically interesting compounds.

The first data set consists of the mass spectra of 11 mono- and sesquiterpenes [9]. These are isoprene (1), myrcene (2), *p*-cymene (3),  $\beta$ -pinene (4), camphene (5), limonene (6),  $\alpha$ -cedrene (7), caryophyllene (8),  $\beta$ -selinene (9), santene (10),  $\delta$ -cadinene (11). Figure 5A shows the plot of the loadings of the first two vectors of the factor weight matrix. These two factors encompass 97% of the total variance. Two clusters of compounds can be seen; only compound 8 seems to lie in between. It turns out that one cluster consists of the monoterpenes and isoprene; the second are of the sesquiterpenes. Compound 8 (caryophyllene) should therefore belong to the second cluster (see below). Because the first factor encounters 94% of the total variance, there is clearly one main factor, i.e., there is one main underlying fragmentation pattern.

The nonlinear multidimensional scaling configuration of these data in two dimensions shows the separation of the two clusters very clearly (Fig. 5B). The very similar fragmentation pattern of isoprene and the monoterpenes is reflected by their close neighborhood within the cluster. The one-dimen-

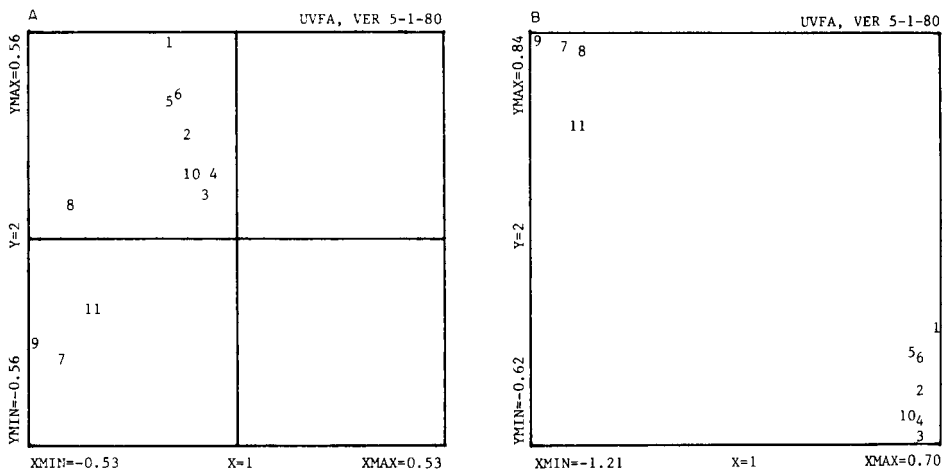


Fig. 5. Terpene mass spectral data: A, plot of the loadings of the first two factors; B, two-dimensional nonlinear projection.

sional multidimensional scaling of the mass spectra corroborates that there is mainly one underlying fragmentation pattern. The stress of the one-dimensional projection is almost as low as for two dimensions (0.0031 and 0.008 respectively) and thus the intrinsic dimensionality is most likely one. (Kruskal [5] considers a stress of 0.1–0.05 as “satisfactory” and below 0.05 as “impressive”; he cautions, however, that a low stress is only a necessary criteria and that a meaningful interpretation of the configuration is most important.)

In the second example, the data base consists of a distance matrix  $D = (d_{ij})$   $i, j = 1, \dots, 13$  of another set of 13 terpene components. These are: isoprene, four monoterpenes (myrcene, menthol, camphene, umbellulone), four sesquiterpenes (bisabolene,  $\alpha$ -cadinol, eudesmol, partheniol), three diterpenes (dextropimaric acid, phyllocladene, roylleanone) and one triterpene ( $\beta$ -amyrin) [10]. The distance measure  $d_{ij}$  is the minimum chemical distance [11] between the compounds  $i$  and  $j$ . It indicates the constitutional similarity between two compounds. To perform a principal component analysis, a covariance matrix  $C$  is generated from the distance matrix [12]:  $d_{ij} = 2(1 - C_{ij})^{1/2}$ . Again, two linear factors (85.8 and 13.1% partial variances) are obtained, and a plot of their loadings (Fig. 6A) shows that there are no particular clusters of compounds. For further interpretation the two-dimensional nonlinear projection (Fig. 6B) is examined. The compounds are now clustered according to whether they are mono-, sesqui-, di- or tri-terpenes. Again the stress for a one-dimensional projection is almost as low as for two dimensions (0.0087 and 0.00859, respectively) which suggests that all these compounds consist of one major structural element, the isoprene unit. This is indicated by the ordering of the compounds along the one-dimensional axis according to their number of isoprene units.



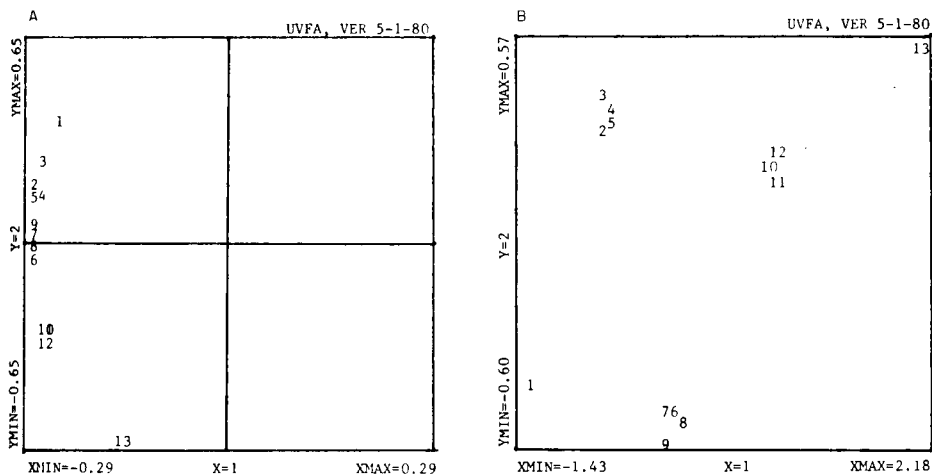


Fig. 6. Minimum chemical distance data of 13 terpenes: A, factor loadings; B, two-dimensional projection.

The third example corroborates some results of a linear factor analysis (principal component analysis) reported by Cramer for a data set of ten physical parameters of 44 organic compounds [13]. Cramer obtained two linear factors with 75.5% and 21% partial variance. A nonlinear projection of the compounds in a two-dimensional space shows no distinct clusters of the compounds (Fig. 7A). Because the data points do not lie along a line

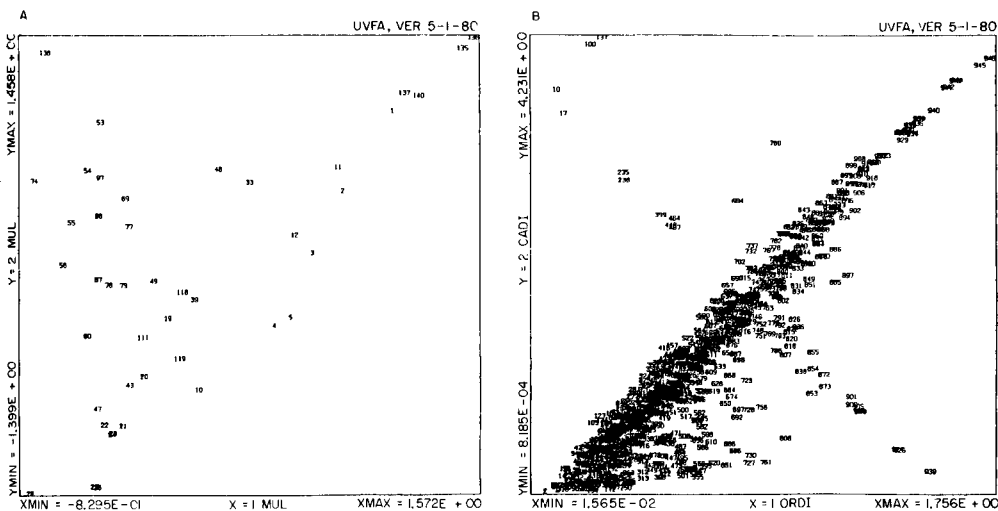


Fig. 7. Data on 10 physical parameters for 44 compounds: A, two-dimensional nonlinear projection; B, original versus calculated distances of the one-dimensional optimum configuration.

but are spread almost equally in both directions, the intrinsic dimensionality seems to be two. An attempt to project the data in a one-dimensional space results in a very interesting pattern of a plot of the calculated distances  $\hat{d}_{ij}$  versus the original  $d_{ij}$  (see above) of the configuration (Fig. 7B). Although most of the distance points lie close to the diagonal, which indicates that most of the compounds fit fairly well in one dimension, some of them almost lie along an axis perpendicular to the diagonal. This probably indicates that there is one major and one minor nonlinear factor in the two-dimensional space.

### Conclusion

The examples show that the combination of principal component analysis with nonlinear least-squares projection techniques is a powerful tool for the determination of the intrinsic dimensionality and the interpretation of the factors. The underlying variables factor analysis program, UVFA, provides a convenient means for complete factor and nonlinear projection analysis of any data set.

We are very grateful for many discussions with Robert Gerlach, and for financial support from the German Academic Exchange Service and the Office of Naval Research.

### REFERENCES

- 1 B. R. Kowalski (Ed.), *Chemometrics: Theory and Application*, Vol. 52, ACS Symposium Series, American Chemical Society, Washington, D.C., 1977.
- 2 K. Überla, *Faktorenanalyse*, Springer-Verlag, Berlin, 1977.
- 3 R. W. Gerlach, B. R. Kowalski and H. Wold, *Anal. Chim. Acta*, 112 (1977) 417; R. W. Gerlach, *Multivariate Methods in Chemistry*, Ph.D. Thesis, University of Washington, Seattle, 1980.
- 4 E. R. Malinowski and D. G. Howery, *Factor Analysis in Chemistry*, Wiley-Interscience, New York, 1980.
- 5 J. B. Kruskal, *Psychometrika*, 29 (1964) 1; 115.
- 6 R. N. Shepard and J. D. Carroll, in P. R. Krishnaiah (Ed.), *International Symposium of Multivariate Analysis*, Academic Press, New York, 1966, p. 561.
- 7 H. C. Andrews, *Introduction to Mathematical Techniques in Pattern Recognition*, Wiley-Interscience, New York, 1972.
- 8 D. L. Duewer, J. R. Koskinen, B. R. Kowalski, *Documentation for ARTHUR*, Version 1-8-75, Chemometrics Society Report No. 2, Seattle, 1975.
- 9 E. Stenhagen, S. Abrahamsson and F. W. McLafferty (Eds.), *Atlas of Mass Spectral Data*, J. Wiley, New York, 1969.
- 10 J. B. Hendrickson, *The Molecules of Nature*, W. A. Benjamin, New York, 1965.
- 11 C. Jochum, J. Gasteiger and I. Ugi, *Angew. Chem. Int. Ed. Engl.*, 19 (1980) 495.
- 12 A. P. M. Coxon and C. L. Jones, in C. A. O'Muircheartaigh and C. Payne (Eds.), *Exploring Data Structures*, Vol. 1, J. Wiley, London, 1977, p. 174.
- 13 R. D. Cramer III, *J. Am. Chem. Soc.*, 102 (1980) 1837; 1849.

## MICROCOMPUTERS IN ELECTROCHEMICAL TRACE ELEMENTAL ANALYSIS

LARS KRYGER

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

(Received 23rd January 1981)

### SUMMARY

The advantages of incorporating digital computers in devices for electrochemical trace elemental analysis are discussed on the basis of examples from the literature. Apart from easing data interpretation, the digital computer may control and monitor electrode processes at a high interaction frequency. For elements undergoing reversible electrode reactions, the resolution and sensitivity of electrochemical techniques can be improved by such rapid real-time interaction. Moreover, for quasi-reversible systems, the application of computerized broad-band potential excitation and simultaneous monitoring of broad-band relaxation signals, permits rapid extraction of analytical information from data which represent the electrode process more accurately than do traditional single-frequency a.c. polarographic data.

In many branches of science, access to reliable and fast methods of determining trace elements is of increasing importance. As the demands for improved analytical performance (resolution, sensitivity, accuracy, precision, etc.) in trace elemental techniques are steadily increasing, research within the field is intense.

Electrochemical techniques of trace elemental analysis are particularly well suited for those elements which can be reduced reversibly at a mercury electrode to form dilute amalgams (e.g., Zn, Cd, Pb, Cu, Ag), and a wealth of reports on the optimization of electrochemical techniques for such elements is available. Optimization of the performance of an electroanalytical technique requires that the chemical as well as the instrumental aspects of the procedure be considered.

Some of the instrumental innovations recently reported involve the incorporation of digital computers in electroanalytical devices. It is the aim of this paper, through examples, to show that the analytical performance of electrochemical techniques is often related to the time scale at which the electrode process can be observed and controlled and that the incorporation of a fast digital computer in electroanalytical devices may improve the analytical performance with respect to those elements which undergo fast electrode processes, i.e., essentially those elements that do not behave completely irreversibly.

*Computer-assisted and computer-dependent techniques*

As in several other fields of analytical chemistry, the application of computers and, more recently, microcomputers in the electrochemical determination of trace elements has become commonplace. Numerous reports on the development and application of computerized electroanalytical methods are currently being published, and a new generation of computerized electroanalytical devices is commercially available.

There are two distinct types of task that can be handled by computerized electroanalytical instruments. The philosophy behind the first category is to make the computer assist the experimenter in obtaining the maximum amount of useful information from the data already acquired, and process and present the data in a convenient form. Such tasks, some of which are listed in Table 1, may sometimes conveniently be carried out in real-time to save computer memory, but, in principle, real-time operation is not crucial and the data processing can be done off-line [1–5]. These techniques may be classified as computer-assisted. For the second category, rapid I/O and real-time control is crucial. Such techniques, which are the main object of the present discussion, exploit the ability of the computer to acquire data at high rates, typically 10–50 kHz, to evaluate the data and to exert a certain amount of control based on this evaluation in real time. By monitoring and controlling the progress of electrode processes in real time, analytical improvements with respect to resolution and sensitivity can be achieved. Moreover, broad-band a.c. polarographic characterization of electrode processes is facilitated, and analytical conclusions based on knowledge of the true nature of the electrode process can be drawn with a turn-round time acceptable for routine work.

While the computer-dependent techniques of the second category are very demanding with respect to real-time accuracy, none of the techniques is particularly demanding with respect to computer size. This means that microcomputers, floppy discs and cassette recorders are suitable components for almost any computerized electroanalytical device.

TABLE 1

Examples of data-processing tasks in electroanalytical chemistry conveniently carried out by computer

Type of data processing	Comment	Ref.
Smoothing	FFT approach	1
Interpolation	FFT approach	2
Differentiation	FFT approach	3
Subtraction of saved voltammograms	Blank correction	4
	In stripping analysis: capacitive background elimination	5
Peak location and integration		4, 5
Conversion of units	In voltammetry, e.g., $\mu\text{A}$ to ppb.	4

Some of the computerized techniques mentioned in Table 1 are implemented on commercially available instruments. The computer-dependent techniques to be dealt with below are so far available only to those who can do some interfacing and programming themselves.

### *The reversible electrode reaction*

From the analytical point of view the preferable electrode process is the simple, diffusion-controlled (reversible) reaction, characterized by a fast heterogeneous electron transfer. Electroanalytical techniques for trace elements based on a reversible electrode reaction are characterized by high sensitivities, and for truly reversible systems, it is probably fair to say that electrochemical analysis is the method of choice. When, however, the heterogeneous electron transfer rate is sufficiently low, i.e., the electrode process is kinetically controlled, small variations of the concentrations of, for example, adsorbing agents from sample to sample may have serious effects on the reproducibility; while diffusion coefficients do not depend significantly on the concentrations of adsorbing agents, heterogeneous rate constants often exhibit a marked dependence on the concentrations of such agents [6, 7]. It is therefore important that the possibility of kinetic control is realized and that any deviations from reversible behaviour of the electrode reaction can be detected. If reversibility is tacitly assumed, an unanticipated complex electrode reaction may give rise to irreproducible results, which are difficult to interpret.

By taking proper chemical steps in the preparation of samples for electrochemical analysis, reversibility can sometimes be achieved. Furthermore, by a proper choice of solvent and carrier electrolyte, and by the addition of suitable complexing agents, the resolution of electroanalytical techniques can in favourable cases be improved. Such precautions are summarized in Table 2, column A; column B shows the computerized techniques that may be adapted for the same reasons and which will be discussed below.

### EXAMPLES

Examples are given to illustrate how the application of rapid real-time control/data acquisition in electroanalytical experiments may lead to improved results. It is important to observe that the computerized techniques described here are suitable for reversible, i.e., essentially fast, electrode processes, and for extracting reversible information from quasi-reversible processes, i.e., processes which appear reversible if studied with a.c. methods at sufficiently low perturbation frequencies. The techniques are generally not suitable for completely irreversible electrode processes. Elements which do not behave reversibly in any medium are probably better determined by some non-electrochemical technique.

TABLE 2

Precautions which may improve the analytical performance of electrochemical techniques

Measure of performance	A	B
	General precautions	Instrumental precautions (computerized methods)
Sensitivity	Choice of medium in which electrode reaction is reversible. For stripping analysis: prolonged deposition time	Extraction of reversible response information from broad-band admittance data. For reversible reactions, multiple scanning techniques High a.c. perturbation amplitude
Resolution	Choice of medium with suitable complexing agents	Resolution of overlapped peaks by interrupted sweep voltammetry. Potentiometric stripping analysis. Low a.c. perturbation amplitude
Inter-sample reproducibility	Choice of medium in which electrode reaction is reversible	Extraction of reversible response information from broad-band admittance data

*Example 1: Real-time kinetic correction*

Traditionally, electrochemical investigations are characterized as either kinetic/mechanistic or analytical studies: while complete mechanistic studies usually require that the electrode process be studied by electrochemical techniques such as a.c. polarography over a wide range of frequencies, analytical studies are mostly carried out at a single frequency. The results obtained by Smith and co-workers [6–13] give good reasons to believe that the performance of electrochemical techniques for trace elemental determinations could be substantially improved if the distinction between kinetic/mechanistic and analytical work was abandoned and if it was generally recognized that complex electrode reactions are very likely so that analytical experiments could be designed accordingly. Nevertheless, the tendency to assume simple reversible electrode reactions for analytical work is understandable, considering that a complete kinetic characterization over several decades of the potential perturbation frequency with conventional electrochemical instrumentation is extremely tedious. By exploiting computerized excitation and acquisition of broad-band faradaic admittance spectra, as proposed by Smith and co-workers, monitoring of the status of the electrode process (e.g., deviation from reversibility) can be carried out in real time. Moreover, for the frequently encountered quasi-reversible process, extraction of the reversible response is possible. Hence normalized information, corrected for the effect of, for example, adsorbing agents, is available.

The computerized generation of broad-band excitation waveforms and acquisition of the faradaic admittance spectra rely on the application of the fast Fourier transform (FFT) and its inverse (FFT<sup>-1</sup>). To characterize the

electrode process, it is often useful to record the frequency domain cell admittance  $A(\omega)$  as a function of frequency  $\omega$ . For small amplitude potential perturbations  $e(t)$  superimposed on a d.c. potential, the admittance spectrum  $A(\omega)$  can essentially be expressed by

$$A(\omega) = I(\omega) E^*(\omega) / E(\omega) E^*(\omega) \quad (1)$$

where  $E(\omega)$  and  $I(\omega)$  denote the Fourier spectra of the applied potential and the observed current-response waveforms, respectively, and where the asterisk indicates complex conjugation. Knowledge of the functions  $A(\omega)$  and  $\cot \Phi(\omega)$  (where  $\Phi(\omega)$  at any frequency  $\omega$  denotes the phase angle of the fundamental harmonic alternating current relative to the applied alternating potential) and the variation of these functions with d.c. potential is normally the information required for kinetic-mechanistic studies.

The reader is referred to the original papers for a complete discussion of the interpretation of broad-band faradaic admittance data and evaluation of the kinetic parameters [7, 8].

For this discussion, however, it is important to give some consideration to the generation of the broad-band excitation waveform and the data-acquisition scheme, because this provides a good illustration of how computerized techniques can be exploited to obtain analytical information otherwise almost inaccessible.

Figure 1 (left half) shows the digital synthesis of a broad-band potential excitation waveform and its conversion to a time series of analog voltages which, when superimposed on a d.c. potential ramp, comprises the excitation signal applied to the electrochemical cell. The computer arrays (a) and (b) are initialized with the phase and amplitude content of the Fourier spectrum of the desired signal. This information is then transformed from polar to complex representation in array (c) and, by an inverse FFT, the digitized perturbation waveform  $e(t)$  is computed and stored in array (d). Periodic potential perturbations representing a wide range of frequencies can now be applied to the cell by outputting the time series  $e(t)$  to the potentiostat using a D/A converter (the output of which is smoothed by a low pass filter) and the programmable clock, PCLOCK. In this manner, the perturbation signal can be "designed" to fulfil various requirements, such that the frequency content adequately covers the range of interest. Moreover, in this way, the possibility of selecting a limited series of perturbation frequencies ensures that the FFT operations conserve the spectral information accurately. The right half of Fig. 1 shows the acquisition of the potential and current vs. time series,  $e(t)$  and  $i(t)$ . Fourier transformation yields the spectra  $E(\omega)$  and  $I(\omega)$  necessary to compute  $A(\omega)$ , according to Eqn. (1), and the phase angle cotangent of the fundamental harmonic as a function of perturbation frequency  $\omega$ .

The computerized strategy outlined here allows the acquisition and evaluation of broad-band data through the entire d.c. range normally of interest within a few minutes. Extraction of analytical information from

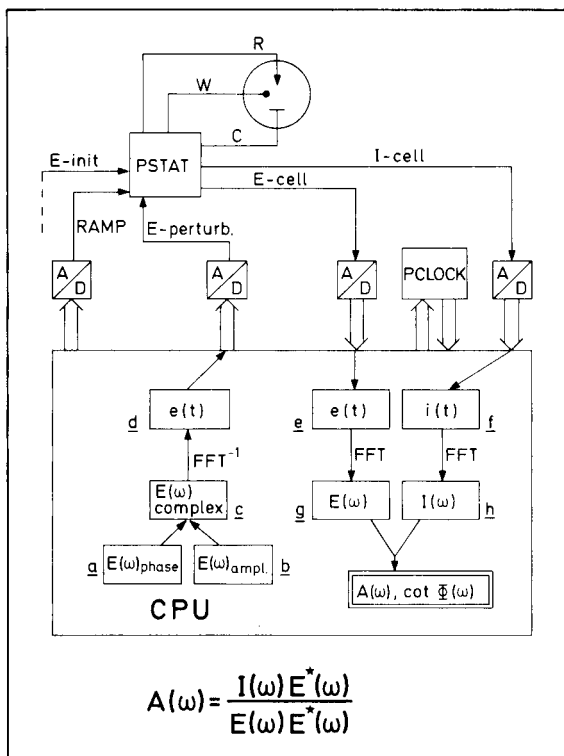


Fig. 1. Computerized generation and acquisition of broad-band waveforms. Left half: design and synthesis of broad-band potential excitation waveform: (a) phase array; (b) amplitude array; (c) complex number array representing (a) and (b); (d) time series of potential perturbations. Right half: Acquisition of broad-band waveforms: (e) sampled potential vs. time series; (f) sampled current vs. time series; (g) Fourier spectrum of potential waveform; (h) Fourier spectrum of current waveform.

FFT, fast Fourier transform;  $\text{FFT}^{-1}$ , inverse fast Fourier transform; PSTAT, potentiostat; PCLOCK, programmable clock pulse generator.

data representing the true mechanistic nature of the system is thus possible on a time scale acceptable for routine analytical work. For the frequently encountered quasi-reversible electrode process, Schwall et al. [7] have shown that it is in fact possible to apply a correction for the kinetic effect and compute the reversible a.c. polarogram at a mercury electrode of cadmium(II) in zinc sulphate for the case where an adsorbing agent, n-butanol, substantially suppresses the value of the heterogeneous charge transfer rate constant.

It is important to notice that the frequency range which can be studied, depends on the frequency at which the computerized device interacts with the experiment. A high interaction frequency therefore increases the probability of assigning the correct mechanism to the electrode process. It is



reasonable to believe that further research within this field of electrochemistry may lead to the development of computerized techniques where the often frustrating kinetic complications may be corrected for automatically.

*Example 2: Resolution enhancement of overlapped voltammetric peaks by the interrupted potential sweep technique*

Although voltammetric techniques can with some justification be regarded as multi-element techniques, a major problem in voltammetric determinations of trace elements is the low resolution of voltammograms. Voltammetric analysis of samples containing several substances with similar half-wave potentials is frequently complicated by overlapped peaks. Various measures, such as the addition of suitable complexing agents, or lowering the potential scan rate, can be taken to enhance the resolution of voltammograms, but frequently these precautions are not sufficiently effective. Perone and co-workers [14, 15] have described a computerized voltammetric mode of operation, the interrupted potential-sweep technique, which may significantly improve the resolution of voltammograms obtained at a stationary electrode.

Figure 2A illustrates the overlap problem occurring when a linear potential ramp (Fig. 2B) is applied to a sample of two reducible species 1 and 2 with similar half-wave potentials: quantitative information about species 1 is readily obtained even at low values of the peak separation. However, because

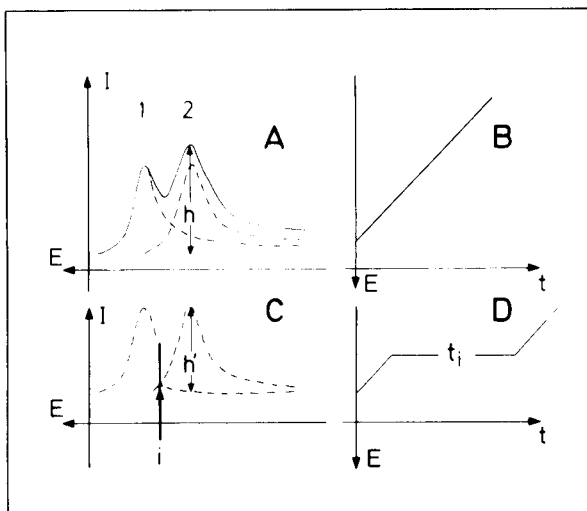


Fig. 2. Interrupted potential-sweep technique applied to two reducible species 1 and 2. A, Overlapped peaks, when there is no interrupt of the potential scan (B). C, Resolution enhancement when the potential scan (D) is interrupted after the reduction peak of 1 has been observed.

peak 1 has a long tail with an appreciable value when the peak potential 2 is reached, the net peak height of 2 is difficult to estimate on the basis of the observed height,  $h$ , particularly when the concentration of 1 is large compared to that of 2. If, however, the potential sweep is interrupted and held for a period  $t_i$  at a suitable potential between the two peaks (Fig. 2D), the diffusion current arising from the reduction of 1 decreases significantly before the scan across 2 is resumed. This happens because, during the interrupt, the potential of which is indicated by  $i$  in Fig. 2C, species 1 is depleted in the vicinity of the electrode surface. The experiment requires that the cell current be monitored and the potential controlled in real time, so that the potential sweep can be interrupted at a potential based on the observation of the first reduction peak and, possibly on approximate knowledge of the composition of the mixture and the peak potentials obtained from a preliminary experiment. Jones and Perone [15] have shown that (100:1) mixtures of thallium(I) and lead(II) ( $5 \times 10^{-4}$  M and  $5 \times 10^{-6}$  M) in 1 M sodium hydroxide can be quantitatively resolved by using the interrupted sweep technique with derivative stationary electrode polarography. By conventional voltage scanning, the lead(II) reduction peak for this solution cannot be detected. It is important to notice that resolution enhancement is obtained by the interrupted sweep technique only when the vicinity of the electrode can be depleted of the species first reduced, i.e., this species must undergo a reversible reduction.

*Example 3: Sensitivity enhancement in computerized potentiometric stripping analysis*

For the electrochemical determination of elements at the ppb/sub-ppb level, stripping methods are almost invariably used. These techniques are particularly sensitive because they involve an electrolytic preconcentration step where analytes are deposited on or dissolved in a suitable working electrode. Until recently the redissolution of such deposits was mostly studied by voltammetric or modified voltammetric techniques to obtain quantitative measures of the analyte concentrations. With voltammetric techniques, the analytical response corresponding to the redissolution of a component at low concentration is a weak current, hence to ensure satisfactory response for a given amount of preconcentrated analyte the properties of the amplification system should be optimized.

Recently, however, a potentiometric mode of detecting the redissolution of preconcentrated analytes has proved feasible for the determination of a wide range of elements [16, 17]. Rather than forcing the redissolution of deposited analytes by scanning the potential, some dissolved oxidizing (or reducing agent) causes the deposit to redissolve while potentiostatic control is abandoned. If the redissolution process is diffusion-controlled, a high concentration region of newly dissolved analyte is formed in the vicinity of the electrode during the stripping [18]. As long as any deposit remains on the electrode, the predominant species at the electrode/solution interface

are therefore the reduced and oxidized forms of the analyte. Hence, the potential vs. time behaviour of the working electrode exhibits a characteristic plateau at the stripping potential of any component, the lengths of the plateaux being a measure of concentrations. In potentiometric stripping analysis (p.s.a.), the analytical signal is time, and so computerized data acquisition is particularly favourable. With a sufficiently high data-acquisition frequency, even transient stripping signals arising from low concentrations of electroactive analytes can be recorded. Hence, whereas voltammetric analysis for trace elements puts requirements on the amplification of currents, the quantity to optimize for high sensitivity in computerized p.s.a. is the data-acquisition rate [19].

Figure 3A shows the potential versus time behaviour of a mercury working electrode during the stripping of amounts of cadmium and lead preplated during time  $t_p$ . The lengths of the plateaux  $a_1$  and  $b_1$  are proportional to the bulk concentrations of the two metals. Figure 3A also illustrates the observation [20] that a considerable fraction of the newly stripped material may be recovered and stripped in subsequent cycles if the working electrode is reset to potentiostatic plating conditions, when a pre-selected limiting potential,  $E_a$ , is reached and kept at the plating potential for a short period  $t'_p$ , even when  $t'_p \ll t_p$ . Hence multiple stripping of material preconcentrated during one major plating period is feasible. If the stripping potentiograms are accumulated in the computer memory, a considerable signal enhancement can be achieved. It is important here that real-time control can be exerted based on evaluations of the acquired potential data: if the electrode potential is allowed to drop beyond the preset limit, the extra time involved allows more material to escape the electrode by diffusion, i.e., the recovery between

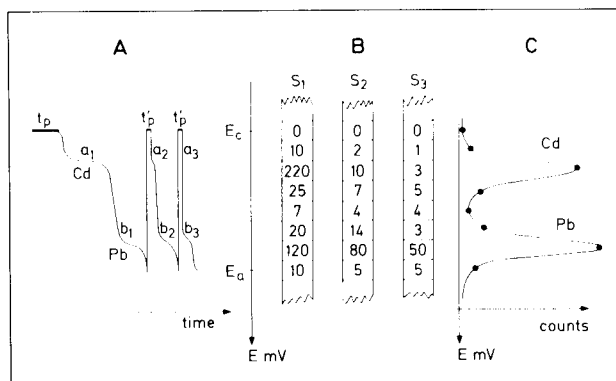


Fig. 3. Computerized potentiometric stripping analysis. A, Potential vs. time behaviour of working electrode during stripping of cadmium and lead following a major pre-electrolysis period  $t_p$ ; multiple replating/stripping cycles with short plating periods  $t'_p$  are also shown. B, Clock pulses accumulated in computer memory during each of the three cycles shown in A. C, Accumulated potential distribution (multichannel potentiogram).

scans and therefore the sensitivity decrease. Figure 3A indicates that the recovery of lead (plateaux  $b_1, b_2, b_3$ ) between successive scans is higher than that of cadmium (plateaux  $a_1, a_2, a_3$ ). This is because the time spent at the stripping potential of lead is available for newly stripped cadmium to escape the electrode by diffusion. By modifying the experiment so that the interval  $E_c - E_a$  is narrow (of the order 50 mV) and moves slowly across the range of interest as a function of scan number, the signal enhancement is improved, and the dependence of the cadmium signal on the lead concentration vanishes [21].

In p.s.a., the independent variable is the potential, while the time spent at a given potential is the quantity to be measured. The simplest mode of data acquisition involves constant monitoring of the potential and accumulation of the number of clock ticks associated with each small potential interval in a data buffer. The address of the location where clock tick accumulation takes place is proportional to the potential reading. Figure 3B illustrates the data which are accumulated and added in real-time to a buffer which is initially reset to all zeroes.  $S_1, S_2$  and  $S_3$  are the counts obtained during the first, second and third cycles, respectively. Figure 3C shows the potential distribution obtained after 3 cycles.

In Fig. 4 the impact on the sensitivity of the data acquisition rate is illustrated: the potential distributions A and B are obtained for a  $10^{-7}$  M cadmium(II) solution under identical conditions, except that in Fig. 4A the data are obtained with a spacing of 3 ms, while in Fig. 4B the spacing is  $750 \mu\text{s}$ .

Maximizing the data acquisition rate and using multiple scanning across small potential intervals [21] leads to high sensitivities for the elements which exhibit reversible redissolution. Figure 5 shows the detection of  $5 \times 10^{-10}$  M cadmium(II) after 60 s of pre-electrolysis using this approach. With a non-computerized approach, a signal of similar quality would require a considerably longer plating period.

## CONCLUSIONS

It appears to be generally accepted that substantial advantages can be gained by incorporating microcomputers in electroanalytical devices. So far the number of computer applications for pure data processing and display purposes clearly exceeds the number of computer-dependent methods where the unique properties of the computer in acquiring data, exerting control, and taking decisions at high rates are exploited to yield improved analytical data. However, further research within this second category of applications is in progress at several laboratories, and the exploitation of new computer generations, capable of I/O at and below the microsecond level, is likely to produce improved results.

As previously pointed out, elements which undergo totally irreversible electrode processes are best determined by non-electrochemical techniques.

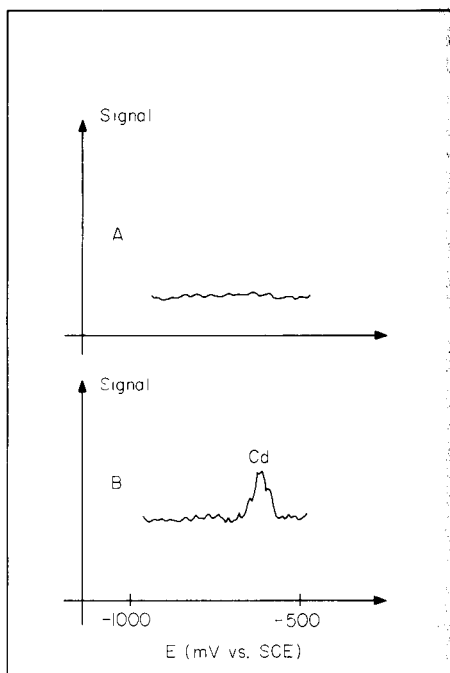
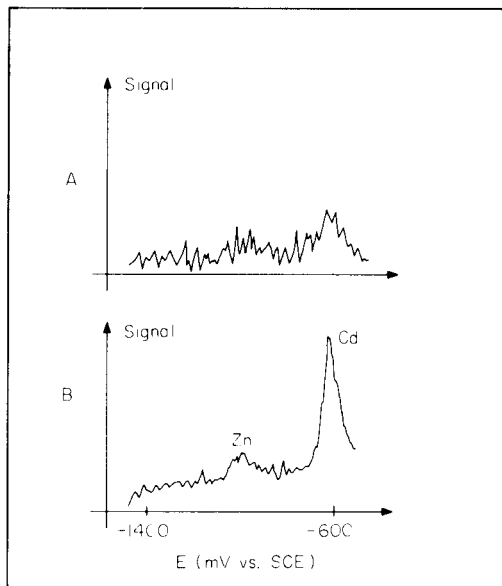


Fig. 4. Dependence of sensitivity in computerized potentiometric stripping analysis on data-acquisition rate for  $10^{-7}$  M cadmium(II) in 0.5 M acetate buffer, pH 4.5; oxidizing agent  $2 \times 10^{-4}$  M mercury(II). A, 3 ms/datum; B, 750  $\mu$ s/datum.

Fig. 5. Detection of  $5 \times 10^{-10}$  M cadmium(II) using 60-s plating by computerized potentiometric stripping analysis in 0.5 M acetate buffer pH 4.5 with  $2 \times 10^{-4}$  M mercury(II). A, Potentiogram obtained from the blank medium; B, potentiogram obtained after the addition of  $5 \times 10^{-10}$  M cadmium(II) to the blank.

Future research on computerized electrochemical methods for trace elemental analysis should therefore concentrate on elements which do not belong to this class.

Grants from the Danish Natural Science Research Council (511-20682) and from Aarhus University which enabled the author to present this paper are gratefully acknowledged.

#### REFERENCES

- 1 J. W. Hayes, D. E. Glover, D. E. Smith and M. W. Overton, *Anal. Chem.*, 45 (1973) 277.
- 2 R. J. O'Halloran and D. E. Smith, *Anal. Chem.*, 50 (1978) 1391.
- 3 R. de Levie, S. Sarangapani, P. Czekaj and G. Benke, *Anal. Chem.*, 50 (1978) 110.
- 4 Princeton Applied Research, Model 384-1 Polarographic Analyzer System, users manuals.
- 5 L. Kryger and D. Jagner, *Anal. Chim. Acta*, 78 (1975) 251.

- 6 D. E. Smith, *Anal. Chem.*, 48 (1976) 517A.
- 7 R. J. Schwall, A. M. Bond and D. E. Smith, *J. Electroanal. Chem.*, 85 (1977) 217.
- 8 R. J. Schwall, A. M. Bond and D. E. Smith, *Anal. Chem.*, 49 (1977) 1805.
- 9 R. J. Schwall, A. M. Bond and D. E. Smith, *Anal. Chem.*, 49 (1977) 1073.
- 10 A. M. Bond, R. J. Schwall and D. E. Smith, *J. Electroanal. Chem.*, 85 (1977) 231.
- 11 R. J. Schwall, A. M. Bond, R. J. Lloyd, J. G. Larsen and D. E. Smith, *Anal. Chem.*, 49 (1977) 1797.
- 12 D. E. Smith, *Anal. Chem.*, 48 (1976) 221A.
- 13 S. C. Creason, J. W. Hayes and D. E. Smith, *J. Electroanal. Chem.*, 47 (1973) 9.
- 14 S. P. Perone, D. O. Jones and W. F. Gutknecht, *Anal. Chem.*, 41 (1969) 1154.
- 15 D. O. Jones and S. P. Perone, *Anal. Chem.*, 42 (1970) 1151.
- 16 D. Jagner and A. Graneli, *Anal. Chim. Acta*, 83 (1976) 19.
- 17 D. Jagner, *Anal. Chem.*, 50 (1978) 1924.
- 18 J. K. Christensen and L. Kryger, *Anal. Chim. Acta*, 118 (1980) 53.
- 19 H. J. Skov and L. Kryger, *Anal. Chim. Acta*, 122 (1980) 179.
- 20 J. Mortensen, E. Ouziel, H. J. Skov and L. Kryger, *Anal. Chim. Acta*, 112 (1979) 297.
- 21 L. Kryger, *Anal. Chim. Acta*, 120 (1980) 19.

## COMPUTER-ASSISTED STRUCTURE—CARCINOGENICITY STUDIES ON POLYCYCLIC AROMATIC HYDROCARBONS BY PATTERN RECOGNITION METHODS

YOSHIKATSU MIYASHITA, TOMOKO SEKI, YOSHIMASA TAKAHASHI,  
SHIN-ICHI DAIBA, YUICHIRO TANAKA, YASUHIKO YOTSUI\*\*, HIDETSUGU ABE  
and SHIN-ICHI SASAKI\*

*School of Materials Science, Toyohashi University of Technology, Tempaku, Toyohashi,  
Aichi 440 (Japan)*

(Received 23rd January 1981)

### SUMMARY

Pattern recognition methods are applied to the study of structure—carcinogenicity relationships in 25 representative polycyclic aromatic hydrocarbons (PAHs). On the basis of presumed metabolic transformation, a variety of reactivity indices taken from simple Hückel molecular orbital theory for not only parent PAH but also later metabolites are used to investigate the carcinogenic process. In order to display the 12-dimensional molecular descriptor space, a Karhunen—Loève plot in two-dimensional space is employed; 92.1% of the variance is retained. The data structure shows asymmetric character. Carcinogens are clustered, whereas non-carcinogens are scattered. Linear discriminant functions for carcinogenicity are developed by using multiple linear regression analysis. The most significant equations suggest the importance of metabolic pathways.

Numerous attempts have been made to explore structure—carcinogenicity relationships in polycyclic aromatic hydrocarbons (PAHs). The K- and L-region theory of Pullman and Pullman [1] and the bay-region theory of Jerina et al. [2, 3] have been applied to this problem. Most of the theoretical models have focused attention on properties of the parent PAHs. Recently, however, the metabolic transformations of benzo[a]pyrene [4, 5] and benz[a]anthracene [6] have been studied. Benzo[a]pyrene and benz[a]-anthracene are metabolically activated and transformed in vivo from pre-carcinogen to ultimate carcinogen. On the basis of this presumed transformation, Smith et al. [7] have qualitatively examined the relationships between carcinogenicity and a variety of reactivity indices taken from simple Hückel molecular orbital theory in 25 unsubstituted PAHs. Pullman [8] has reviewed critically recent discoveries on the metabolic transformations of PAHs.

Pattern recognition methods have been applied to structure—activity studies [9, 10]. These techniques have also been used for structure—carcino-

---

\*\*Present address: Research Institute, Daiichi Seiyaku Co. Ltd., Edogawa-ku, Tokyo 132, Japan.

genicity studies [11–15]. In this report, pattern recognition methods are applied to the study of structure–carcinogenicity relationships using theoretical indices relating to a series of metabolic transformations obtained by Smith et al.

## DATA SET

As a result of a variety of studies, the preliminary path by which benzo[*a*]pyrene is metabolically activated and transformed *in vivo* from precarcinogen to ultimate carcinogen is believed to consist of the stages shown in Fig. 1. Only one stereoisomeric form is illustrated. In step (a) of Fig. 1, benzo[*a*]pyrene is converted to its 7,8-epoxide IIa at the A-region. This is the presumptive initial epoxidation site on the terminal ring of the bay region (called the M-region by Pullman [8]). In step (b), the epoxide IIa is transformed to the 7,8-dihydrodiol IIb. Saturation of the 7,8-bond activates the 9,10-bond, the B-region. This is the site of final epoxidation on the terminal ring of the bay-region (called the N-region by Pullman [8]). Compound IIb is transformed to the 7,8-dihydrodiol-9,10-epoxide III in step (c). In step (d), the diol-epoxide III converts spontaneously to the triol carbonium ion IV. It has been suggested [16] that carbonium ions such as IV act as ultimate carcinogens via electrophilic attack on critical cellular nucleophiles, e.g., DNA. Compounds IIa and IIb are equivalent from the Hückel molecular orbital theory; they have the same theoretical indices.

The carcinogenicity indices of Arcos and Argus [17] and of Jerina et al. [2, 3] for PAHs are shown in Table 1. Fourteen molecular structure descriptors are shown in Table 2. Thus the chemical structures of PAH and its metabolites are represented by a 14-dimensional pattern vector. Table 3 shows which descriptor is associated with a series of metabolic transformations.

## CLUSTERING OF POLYCYCLIC AROMATIC HYDROCARBONS

The preprocessing method is autoscaling to weight all descriptors equally. This method provides zero mean and unit standard deviation for all descrip-

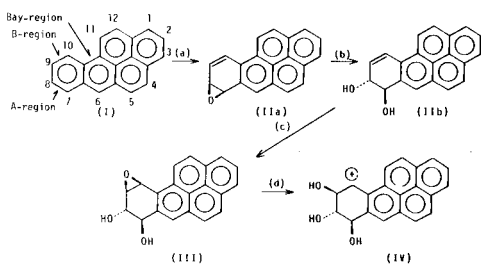


Fig. 1. Metabolic conversion of benzo(*a*)pyrene.



TABLE 1

Carcinogenicity indices for 25 polycyclic aromatic hydrocarbons

Compound	Name	Carcinogenicity index	
		Arcos and Argus [17]	Jerina et al. [2, 3]
1	Naphthalene	0	—
2	Anthracene	0	—
3	Tetracene	0	—
4	Pentacene	0	—
5	Hexacene		?
6	Benz[a]anthracene	5	+
7	Benzo[a]tetracene		—
8	Phenanthrene	0	—
9	Benzo[c]phenanthrene	4	+
10	Chrysene	3	+
11	Benzo[b]chrysene		—
12	Picene	0	—
13	Triphenylene	0	—
14	Benzo[g]chrysene	17	2+
15	Dibenz[a,c]anthracene	3	+
16	Dibenz[a,j]anthracene	4	+
17	Dibenz[a,h]anthracene	26	2+
18	Naphtho[2,3-a]pyrene	27	2+
19	Benzo[a]pyrene	73	4+
20	Benzo[e]pyrene	2	+
21	Dibenzo[a,l]pyrene	33	2+
22	Dibenzo[a,i]pyrene	74	4+
23	Dibenzo[a,e]pyrene	50	3+
24	Dibenzo[a,h]pyrene	70	4+
25	Tribenzo[a,e,i]pyrene	16	2+

tors. Both correlation coefficients between  $P_A$  and  $\Delta E_\pi^{(1)}$ , and  $Q_b$  and  $\Delta E_{deloc}$  are 0.999. Therefore,  $P_A$  and  $Q_b$  are omitted and the remaining twelve-dimensional data are analyzed. The correlation matrix for 12 descriptors is shown in Table 4. A hierarchical clustering method [18] was applied to 25 PAHs. Here, the distance between two clusters is determined by the nearest neighbours in the two clusters. The result is shown as a branching-tree diagram in Fig. 2. It is clear that carcinogenic compounds are clustered, whereas noncarcinogenic compounds are scattered. This result suggests the asymmetric nature of molecular descriptor space. The  $K$ -nearest neighbor classification rule ( $K = 1$ ) was applied to 24 PAHs for carcinogenic data (Arcos and Argus [17]). On the basis of autoscaled 12 structural descriptors, the predictive ability for classifying PAHs as carcinogens or noncarcinogens is 79.2%.

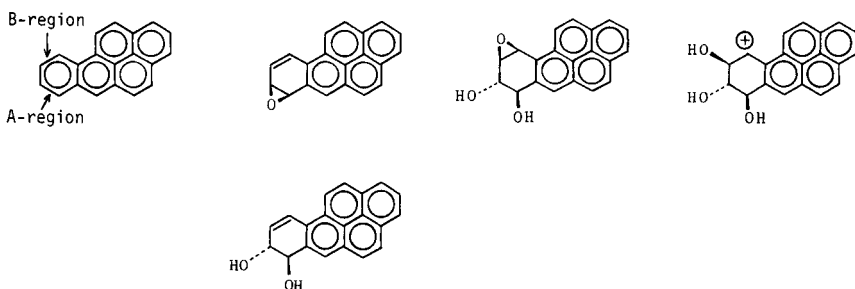
TABLE 2

## Molecular structure descriptors for polycyclic aromatic hydrocarbons

No.	Descriptor	
1	$\ln P$	log of partition coefficient $P$ for parent compound
2	$L_A$	sum of two atomic superdelocalizabilities involved in an $A$ -regional bond for parent compound
3	HOMO	highest occupied molecular orbital energy for parent compound
4	$L_{B'}$	bond superdelocalizability for $A$ -region dihydrodiol form
5	$P_{B'}$	bond order for the $A$ -region dihydrodiol form
6	$\Delta E_{\text{deloc}}$	change in delocalization energy
7	$F_b$	carbonium ion free valence
8	$S_b$	carbonium ion atomic superdelocalizability
9	$\Delta E_{\pi}^{(1)}$	energy loss in going from the parent compound to the $A$ -region epoxide or dihydrodiol
10	$\Delta E_{\pi}^{(2)}$	energy loss in forming the dihydrodiol-epoxide from the $A$ -region dihydrodiol
11	$\Delta E_{\pi}^{(3)}$	energy change in forming the trihydrotriol carbonium ion from the dihydrodiol-epoxide
12	NC	no. of carbon atoms
13	$P_A$	bond order for the $A$ -region parent compound
14	$Q_b$	carbonium ion charge density at the benzylic carbon position

TABLE 3

## Metabolites and descriptors



$\ln P$	NC	$\Delta E_{\pi}^{(1)}$	$L_{B'}$	$\Delta E_{\pi}^{(2)}$	$\Delta E_{\pi}^{(3)}$	$\Delta E_{\text{deloc}}$	$S_b$
$L_A$	$P_A$		$P_{B'}$			$F_b$	$Q_b$
HOMO							

## Display of data structure

The advantage of the display method is that it offers easy understanding of the complicated data structure in a high-dimensional space [19, 20]. If the intrinsic dimensionality of the data is high, the nonlinear mapping method is preferred. However, it should be remembered that the axes given

TABLE 4

Correlation matrix for twelve structure descriptors

	lnP	$L_A$	HOMO	$L_{B'}$	$P_{B'}$	$\Delta E_{\text{deloc}}$	$F_b$	$S_b$	$\Delta E_{\pi}^{(1)}$	$\Delta E_{\pi}^{(2)}$	$\Delta E_{\pi}^{(3)}$	NC
lnP	1.000											
$L_A$	0.590	1.000										
HOMO	-0.790	-0.842	1.000									
$L_{B'}$	0.584	-0.013	-0.374	1.000								
$P_{B'}$	-0.313	0.370	0.014	-0.909	1.000							
$\Delta E_{\text{deloc}}$	0.707	0.125	-0.491	0.981	-0.834	1.000						
$F_b$	-0.434	0.255	0.139	-0.952	0.988	-0.903	1.000					
$S_b$	0.876	0.473	-0.735	0.792	-0.502	0.890	-0.620	1.000				
$\Delta E_{\pi}^{(1)}$	0.154	0.798	-0.582	-0.350	0.600	-0.275	0.533	0.021	1.000			
$\Delta E_{\pi}^{(2)}$	-0.233	0.457	-0.083	-0.859	0.994	-0.771	0.968	-0.415	0.653	1.000		
$\Delta E_{\pi}^{(3)}$	0.474	-0.212	-0.209	0.950	-0.970	0.903	-0.985	0.635	-0.478	-0.947	1.000	
NC	0.922	0.283	-0.574	0.710	-0.557	0.783	-0.641	0.796	-0.170	-0.499	0.655	1.000

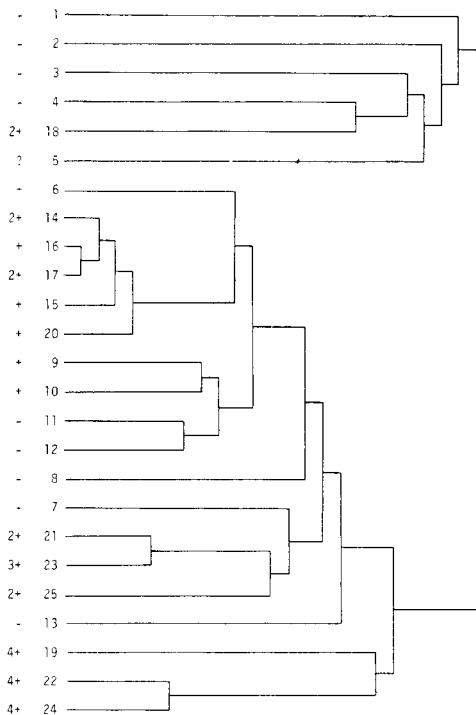


Fig. 2. Branching-tree diagram for 25 PAHs.

by nonlinear transformation have no physical meaning with respect to the original axes.

In contrast, if the intrinsic dimensionality of the data is low, linear mapping is preferred. The Karhunen—Loève transform is one of the linear mapping methods. Each new transformed axis is a linear combination of the descriptors  $x_k$  and is orthogonal to the other axes. The method starts by diagonalizing the covariance matrix  $C$  to obtain eigenvector matrix  $T$  and eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_d$ .

$$T^{-1} C T = \Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdot & \\ 0 & & & \lambda_d \end{pmatrix} \quad (1)$$

with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  and  $T = (t^{(1)}, t^{(2)}, \dots, t^{(d)})$ .

Next, for display purpose in two-dimensional space, the eigenvectors  $t^{(1)}$  and  $t^{(2)}$  associated with the two largest eigenvalues  $\lambda_1$  and  $\lambda_2$  are used for calculating new axes,  $Z_1$  and  $Z_2$ :

$$Z_1 = \sum t_k^{(1)} x_k \text{ and } Z_2 = \sum t_k^{(2)} x_k$$

This K-L plot in two-dimensional space is the best projection that has minimum mean squared error of variance.

The reliability of interpretation using this plot is calculated by the percent variance (%Var) retained by this method:  $\%Var = [(\lambda_1 + \lambda_2) / \sum \lambda_k] \times 100$ . If %Var is near 100, the plot is satisfactory for displaying the data structure.

In order to display the multidimensional data structure, the Karhunen-Loève (K-L) transform is applied to the autoscaled molecular descriptors of PAHs and their metabolites. The K-L plot in two-dimensional space ( $Z_1$ ,  $Z_2$ ) for 25 PAHs is shown in Fig. 3. Since the percent variances retained by the K-L plot in two- and three-dimensional space are 92.1 and 96.7, respectively, the reliability of interpretation of the two-dimensional map is very high. This map is helpful in understanding the multidimensional data structure. It seems that  $Z_1$  and  $Z_2$  axes are related to the length and roundness of PAH, respectively. An asymmetric type of data structure can be observed [21]. As the carcinogenic process for PAHs involves a series of metabolic transformations, it would be reasonable to suppose that carcinogenic compounds and their metabolites have similar properties. The non-carcinogenicity may be due to several factors.

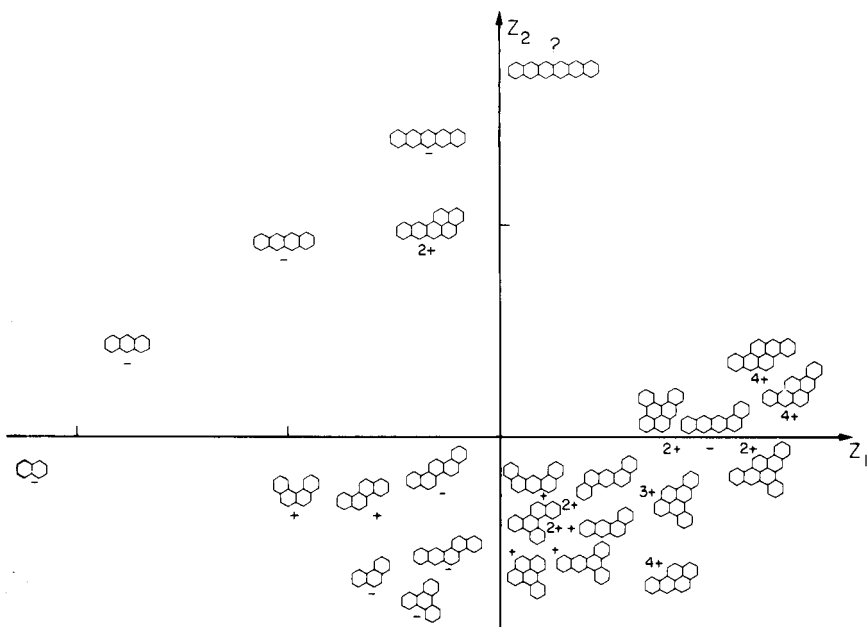


Fig. 3. K-L Plot of 25 PAHs.

## DISCRIMINANT FUNCTION

The display method showed that the characteristics of the parent PAH and later metabolites are critical in determining carcinogenicity in PAHs. It is interesting to correlate the descriptors quantitatively with carcinogenic activity. Linear discriminant functions for carcinogenic activity were developed by using multiple linear regression analysis. Regression equations in which descriptors are highly collinear were eliminated. The following equations were developed for the logarithm of the carcinogenicity indices of Arcos and Argus ( $n = 22$ ).

$$\log A = 9.980L_{B'} - 22.158 \quad (r = 0.794, s = 0.448, F = 34.08) \quad (2)$$

$$\log A = 5.871\Delta E_{\text{deloc}} - 3.376 \quad (r = 0.776, s = 0.464, F = 30.29) \quad (3)$$

$$\log A = -8.847Q_b + 4.498 \quad (r = 0.760, s = 0.478, F = 27.40) \quad (4)$$

$$\log A = 5.537P_A + 10.711L_{B'} - 27.767 \quad (r = 0.805, s = 0.448, F = 17.49) \quad (5)$$

$$\log A = 1.448\Delta E_{\pi}^{(1)} + 10.628L_{B'} - 18.861 \quad (r = 0.805, s = 0.448, F = 17.47) \quad (6)$$

$$\log A = -0.731\text{HOMO} + 9.396L_{B'} - 20.486 \quad (r = 0.801, s = 0.452, F = 17.06) \quad (7)$$

Here,  $n$  represents the number of data points on which the equation is based,  $r$  is the correlation coefficient,  $s$  is the standard deviation, and  $F$  is the  $F$  ratio. Equations (5–7) suggest the importance of the metabolic pathway presumed.

For the carcinogenic PAHs only, the following equations were obtained ( $n = 15$ ).

$$\log A = 7.097L_A - 12.272 \quad (r = 0.778, s = 0.371, F = 19.95) \quad (8)$$

$$\log A = -5.223\text{HOMO} + 3.436 \quad (r = 0.767, s = 0.379, F = 18.61) \quad (9)$$

$$\log A = 5.470\Delta E_{\text{deloc}} - 2.947 \quad (r = 0.708, s = 0.417, F = 13.09) \quad (10)$$

$$\log A = 6.763L_A - 40.207\Delta E_{\pi}^{(2)} - 110.160 \quad (r = 0.896, s = 0.273, F = 24.41) \quad (11)$$

$$\log A = 6.472L_A - 69.451P_{B'} + 51.171 \quad (r = 0.895, s = 0.274, F = 24.23) \quad (12)$$

$$\log A = 6.204L_A - 16.503F_b + 6.416 \quad (r = 0.888, s = 0.283, F = 22.34) \quad (13)$$

It is impressive that the regression equations with two descriptors are highly significant. The correlation coefficients between two descriptors in eqns. (11–13) are  $-0.077$ ,  $-0.149$ , and  $-0.220$ , respectively. The descriptor values and carcinogenicity calculated by eqn. (11) are shown in Table 5.

TABLE 5

Descriptor values and carcinogenicity (Arcos and Argus [17])

No.	$L_A$	$\Delta E_\pi^{(2)}$	log A	
			Obsd.	Calcd. <sup>a</sup>
6	1.848	-2.4504	0.699	0.861
9	1.859	-2.4399	0.602	0.513
10	1.856	-2.4444	0.477	0.674
14	1.870	-2.4494	1.230	0.970
15	1.810	-2.4484	0.477	0.524
16	1.875	-2.4486	0.602	0.971
17	1.848	-2.4495	1.415	0.825
18	2.022	-2.4337	1.431	1.367
19	1.945	-2.4550	1.863	1.702
20	1.813	-2.4466	0.301	0.472
21	1.951	-2.4507	1.519	1.570
22	1.943	-2.4577	1.869	1.797
23	1.908	-2.4519	1.699	1.327
24	1.986	-2.4565	1.845	2.040
25	1.904	-2.4536	1.204	1.369

$$^a \text{Log } A = 6.763L_A - 40.207\Delta E_\pi^{(2)} - 110.160.$$

Linear discriminant functions were developed based on the carcinogenicity indices of Jerina et al. The carcinogenic indices -, +, 2+, 3+, and 4+ are assigned the numerical values 0, 1, 2, 3, and 4, respectively. The following equations were developed ( $n = 24$ ).

$$L = 17.067L_{B'} - 37.970 \quad (r = 0.716, s = 0.976, F = 23.08) \quad (14)$$

$$L = 22.325\Delta E_\pi^{(3)} - 17.055 \quad (r = 0.691, s = 1.010, F = 20.09) \quad (15)$$

$$L = 9.614\Delta E_{\text{deloc}} - 5.550 \quad (r = 0.675, s = 1.031, F = 18.42) \quad (16)$$

$$L = 6.806L_A - 119.179\Delta E_\pi^{(2)} - 303.106 \quad (r = 0.760, s = 0.929, F = 14.37) \quad (17)$$

$$L = 33.421P_A - 247.663P_{B'} + 200.280 \quad (r = 0.760, s = 0.930, F = 14.33) \quad (18)$$

$$L = 8.709\Delta E_\pi^{(1)} - 242.254P_{B'} + 247.914 \quad (r = 0.759, s = 0.931, F = 14.29) \quad (19)$$

For the carcinogenic compounds only, the following equations were obtained ( $n = 15$ ).

$$L = 19.109L_{B'} - 42.335 \quad (r = 0.761, s = 0.783, F = 17.86) \quad (20)$$

$$L = -10.343\text{HOMO} + 6.595 \quad (r = 0.744, s = 0.807, F = 16.09) \quad (21)$$

$$L = 11.611\Delta E_{\text{deloc}} - 6.628 \quad (r = 0.736, s = 0.817, F = 15.36) \quad (22)$$

$$L = 11.931L_A - 107.406\Delta E_{\pi}^{(2)} - 283.610 \quad (r = 0.899, s = 0.550, F = 25.27) \quad (23)$$

$$L = 11.145L_A - 184.581P_B + 146.493 \quad (r = 0.896, s = 0.558, F = 24.41) \quad (24)$$

$$L = 9.760L_A + 21.427\Delta E_{\pi}^{(3)} - 34.431 \quad (r = 0.877, s = 0.604, F = 19.98) \quad (25)$$

The discriminant functions for carcinogenicity of Jerina et al. [2, 3] are very similar to those for carcinogenicity of Arcos and Argus [17].

## DISCUSSION

Cluster analysis and linear mapping are used for understanding the multi-dimensional structures descriptor space. The intrinsic dimensionality of the data is about three. The quantitative analysis to develop discriminant functions shows the importance of the reactivity of the A-region of parent PAHs and the indices for the later metabolites. The result that more significant linear discriminant functions are obtained for only the carcinogenic compounds shows the asymmetric nature of the descriptor data space structure [21]. Smith et al. [7] could find no good explanation for the carcinogenicities of benzo[a]tetracene (Compound 7) and naphtho[2,3-a]pyrene (Compound 18). It can be seen from Table 5 that the calculated carcinogenicity for Compound 18 is in fair agreement with the observed value. This success obtained by using multidescriptors strongly supports the presumed model of metabolic transformations. The prediction of non-carcinogenic compounds in this study also could not explain the non-carcinogenicity of Compound 7. An explanation may be that the model for non-carcinogenicity would require deactivating processes such as the L-region model.

## REFERENCES

- 1 A. Pullman and B. Pullman, *Adv. Cancer Res.*, 3 (1955) 117.
- 2 D. M. Jerina and R. E. Lehr, in V. Ullrich, I. Roots, A. G. Hildebrandt, R. W. Estabrook and A. H. Conney (Eds.), *Microsomes and Drug Oxidations*, Pergamon, Oxford, 1977, 709pp.
- 3 D. M. Jerina, R. Lehr, M. Schaefer-Ridder, H. Yagi, J. M. Karle, D. R. Thakker, A. H. Wood, A. Y. H. Lu, D. Ryan, S. West, W. Levin and A. H. Conney, in H. Hiatt, J. D. Watson and I. Winstin (Eds.), *Origins of Human Cancer*, Cold Spring Harbor, NY, 1977, 639pp.
- 4 P. Sims, P. L. Grover, A. Swaisland, K. Pal and A. Hewer, *Nature*, 252 (1974) 326.
- 5 A. Borgen, H. Darvey, N. Castagnoli, T. T. Crocker, R. E. Rasmussen and I. Y. Wang, *J. Med. Chem.*, 16 (1973) 502.
- 6 A. W. Wood, W. Levin, A. Y. H. Lu, D. Ryan, S. B. West, R. E. Lehr, M. Schaefer-Ridder, D. M. Jerina and A. H. Conney, *Biochem. Biophys. Res. Commun.*, 72 (1976) 680.



- 7 I. A. Smith, G. D. Berger, P. G. Seybold and M. Servé, *Cancer Res.*, 38 (1978) 2968.
- 8 B. Pullman, *Int. J. Quant. Chem.*, 16 (1979) 669.
- 9 G. L. Kirschner and B. R. Kowalski, in E. J. Ariëns (Ed.), *Drug Design*, Vol. 8, Academic, New York, 1979, 73pp.
- 10 A. J. Stuper, W. E. Brügger and P. C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function*, Wiley, New York, 1973.
- 11 B. Nordén, U. Edlund and S. Wold, *Acta Chem. Scand.*, Sect. B, 32 (1978) 602.
- 12 W. J. Dunn III and S. Wold, *J. Med. Chem.*, 21 (1978) 1001.
- 13 P. C. Jurs, J. T. Chou and M. Yuan, *J. Med. Chem.*, 22 (1979) 476.
- 14 J. T. Chou and P. C. Jurs, *J. Med. Chem.*, 22 (1979) 792.
- 15 M. Yuan and P. C. Jurs, *Toxicol. Appl. Pharmacol.*, 52 (1980) 294.
- 16 P. B. Hulbert, *Nature*, 256 (1975) 146.
- 17 J. C. Arcos and M. F. Argus, *Adv. Cancer Res.*, 11 (1968) 305.
- 18 Y. Takahashi, Y. Miyashita, H. Abe, S. Sasaki, Y. Yotsui and M. Sano, *Anal. Chim. Acta*, 122 (1980) 241.
- 19 B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, 94 (1972) 5632; 95 (1973) 686.
- 20 Y. Miyashita, Y. Takahashi, Y. Yotsui, H. Abe and S. Sasaki, *Proceedings of 7th International CODATA Conference*, No. 41. Pergamon, Oxford, 1981, p. 37.
- 21 W. J. Dunn III and S. Wold, *J. Med. Chem.*, 23 (1980) 595.

## APPLICATION OF PATTERN RECOGNITION TO STRUCTURE—ACTIVITY PROBLEMS

### Use of minimal spanning tree

YOSHIKATSU MIYASHITA, YOSHIMASA TAKAHASHI, YASUHIKO YOTSUI\*\*,  
HIDETSUGU ABE and SHIN-ICHI SASAKI\*

*School of Materials Science, Toyohashi University of Technology Tempaku, Toyohashi,  
Aichi, 440 (Japan)*

(Received 23rd January 1981)

#### SUMMARY

A graph-theoretical algorithm based on the minimal spanning tree (MST) is applied to structure–activity problems. The method is helpful in interpreting the results of cluster analysis, and becomes useful by combining with the mapping method that illustrates approximations of a multidimensional data structure. The antibacterial spectra of cephalosporins are analyzed by the MST approach and a linear mapping method. The main diameter obtained by MST gives the representative data set and clarifies the substituent effect on the antibacterial spectra. Relations between the central nervous system activity of benzodiazepine derivatives and their physicochemical parameters are also analyzed by MST and nonlinear mapping methods. These results for cephalosporins and benzodiazepines prove that MST is very useful in understanding the position of compounds in feature space and their activities.

It is very important to interpret multidimensional data structures so that they can be easily visualized. In such cases, two kinds of method, the mapping and clustering methods of pattern recognition, have been used to give approximate representations of high dimensions in a visual space. Applications of the mapping methods in interpreting chemical data have been presented by Kowalski and Bender [1, 2]. Linear and nonlinear display methods can be used to represent multivariate chemical data in two dimensions, thereby allowing the data to be explained in approximately visual status. Clustering methods are perhaps less familiar to chemists, but are becoming more popular. These techniques attempt to group the points in the multidimensional space into (usually) unconnected sets which, it is hoped, will correspond to marked features of the sample. The grouped sets of points are further grouped into larger sets, so that all the points are hierarchically clustered.

Some applications of cluster analysis coupled with nonlinear or linear mapping methods in studies of structure–activity relationships of antibiotics

\*\*Present address: Research Institute, Daiichi Seiyaku Co. Ltd., Edogawa-ku, Tokyo, 132, Japan.

have been reported [3, 4]. The results confirmed that the technique is potentially useful for displaying multidimensional data structures. In the present paper, another approach to visualizing the data structure in a certain multidimensional space is discussed. The approach is based on the graph-theory algorithm termed the minimal spanning tree.

### MINIMAL SPANNING TREE

Zahn [5] has suggested that the graph-theory algorithm based on the minimal spanning tree (MST) is capable of detecting several different clustering structures in arbitrary point sets; the detected clusters can be described in some cases by extensions of the method.

The MST shows a reasonable relation of neighbours in the original data space. Detailed information involved in the data set of interest can thus be obtained. In what follows, an algorithm proposed by Kruskal [6] will be described. Basic definitions of graph theory may be found in Zahn's study [5].

An edge-weighted linear graph is composed of a set of points called nodes and a set of node pairs is called an edge with a number called a weight assigned to each edge. Graphs are helpful in discussing these geometric relationships, and the following example may help in understanding the concepts of the theory. A weighted linear graph with six nodes,  $G$  is illustrated in Fig. 1(a). A path in the graph is a sequence of edges joining two nodes, e.g., (ABC) or (CDAEF). A circuit is a closed path, e.g., (ABCD). A connected graph has paths between any pair of nodes. A tree is a connected graph without circuits. A spanning tree of the connected graph  $G$  is the tree which contains all nodes of  $G$ . If the weight of a tree is defined to be the sum of the weights of its constituent edges, then the spanning tree whose weight is minimum among all the spanning trees of  $G$ , is termed the minimal spanning tree of the graph (Fig. 1b).

The MST searching algorithm proposed by Kruskal [6] will be described for the case of graph  $G$  in Fig. 1. Table 1 shows the process of MST searching. Initially, the edges of the presented graph are arranged in increasing order with respect to the assigned weights and the edges are selected so as not to close a circuit with those previously chosen. After  $(n - 1)$  edges have been

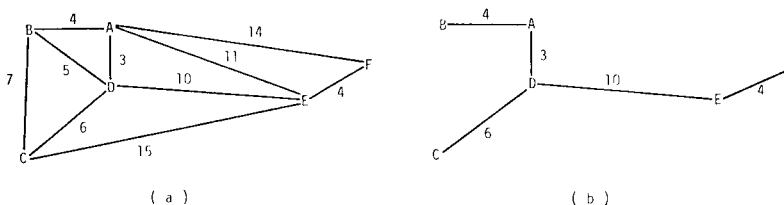


Fig. 1. Graph and minimal spanning tree. (a) Weighted linear graph,  $G$ . (b) Minimal spanning tree.

TABLE 1

A searching process for the minimal spanning tree (MST)

Edge	Weight	Circuit	MST edge
AD	3		o
AB	4		o
EF	4		o
BD	5	(ABDA)	x
CD	6		o
BC	7	(ABCDA)	x
DE	10		o
AE	11	(ADEA)	x
.	.	.	.
.	.	.	.
.	.	.	.

selected, where  $n$  is the number of nodes in the graph, this process is terminated; thus the spanning tree consists of  $(n - 1)$  edges on their node pairs. If it is supposed that the nodes correspond to sample points and the weights of the edges correspond to the distances or similarities between sample points in the data space, the nearest neighbour algorithm in hierarchic cluster analysis could be viewed as an algorithm for finding a minimal spanning tree. The MST approach will be applied to the data set under investigation.

#### DATA SET

Antibacterial spectral data of cephalosporins (I) were obtained from the papers of DeMarinis et al. [7–9]. The chemical structures of fifty-nine 3,7-disubstituted cephalosporins are shown in Table 2. The minimum inhibitory concentration (MIC; mmol l<sup>-1</sup>) data against four gram-negative bacteria (*Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella paratyphi*, *Enterobacter aerogenes*) were used. In order to compare the activities of these compounds with those of clinically available compounds, three cephalosporins (No. 60, 61 and 62; cephaloglycine, cephalothin, and cefazolin, respectively) were included in the data set.

For the benzodiazepine derivatives (II), central nervous system (CNS) activities as given by the  $ED_{50}$  (mmol kg<sup>-1</sup>) values in the inclined screen test for mice [10] were used.

#### ANTIBACTERIAL SPECTRAL DATA ANALYSIS

Zahn's MST approach was applied to the antibacterial spectra of the 62 cephalosporin derivatives specified. In this case, an edge weight on a pair of nodes ( $ij$ ) is the Euclidean distance defined as

TABLE 2

## cephalosporin analogs

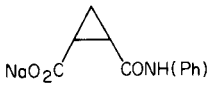
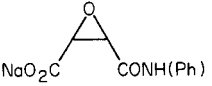
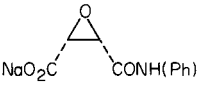
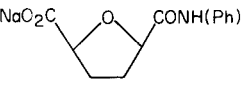
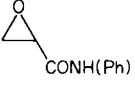

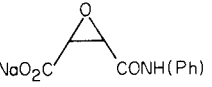
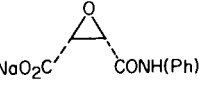
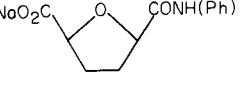
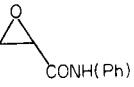
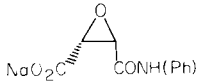
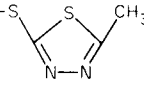
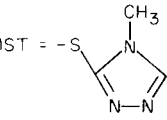
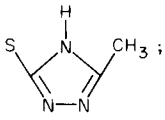
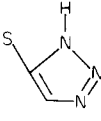
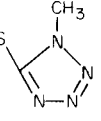
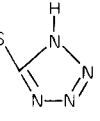
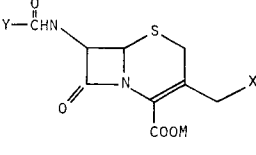
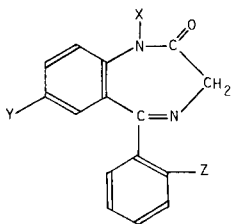
No.	X <sup>a</sup>	Y	M	No.	X <sup>a</sup>	Y	M
1	OAc	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na	46	OAc		Na
2	SMTD	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
3	SMTZ	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
4	S-4MST	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
5	S-5MST	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na	47	OAc		Na
6	SHTZ	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
7	SHTL	CH <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
8	OAc	C <sub>2</sub> H <sub>5</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
9	OAc	n-C <sub>3</sub> H <sub>7</sub> SO <sub>2</sub> CH <sub>2</sub>	H				
0	OAc	n-C <sub>4</sub> H <sub>9</sub> SO <sub>2</sub> CH <sub>2</sub>	H	48	OAc		Na
1	SMTD	C <sub>2</sub> H <sub>5</sub> SO <sub>2</sub> CH <sub>2</sub>	H				
2	SMTD	n-C <sub>3</sub> H <sub>7</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
3	SMTD	n-C <sub>4</sub> H <sub>9</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
4	SMTZ	C <sub>2</sub> H <sub>5</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
5	SMTZ	n-C <sub>3</sub> H <sub>7</sub> SO <sub>2</sub> CH <sub>2</sub>	Na	49	OAc		Na
6	SMTZ	n-C <sub>4</sub> H <sub>9</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
7	OAc	CF <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	H				
8	OAc	NH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
9	OAc	C <sub>6</sub> H <sub>5</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
0	OAc	CH <sub>3</sub> SO <sub>2</sub> CH(CH <sub>3</sub> )	Na	50	OAc		Na
1	SMTZ	CF <sub>3</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
2	SMTZ	NH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
3	SMTZ	C <sub>6</sub> H <sub>5</sub> SO <sub>2</sub> CH <sub>2</sub>	H	51	SMTZ	NaO <sub>2</sub> C(CH <sub>2</sub> ) <sub>3</sub> CONH(Ph)	Na
4	SMTZ	CH <sub>3</sub> SO <sub>2</sub> CH(CH <sub>3</sub> )	Na	52	SMTZ	NaO <sub>2</sub> CCH <sub>2</sub> SCH <sub>2</sub> CONH(Ph)	Na
5	OAc	CNCH <sub>2</sub> SCH <sub>2</sub>	Na	53	SMTZ	NaO <sub>2</sub> CCH <sub>2</sub> OCH <sub>2</sub> CONH(Ph)	Na
6	OAc	CNCH <sub>2</sub> SOCH <sub>2</sub>	Na	54	SMTZ		Na
7	OAc	CNCH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	H				
8	SMTD	CNCH <sub>2</sub> SCH <sub>2</sub>	Na				
9	SMTD	CNCH <sub>2</sub> SOCH <sub>2</sub>	Na				
0	SMTD	CNCH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	H				
1	SMTZ	CNCH <sub>2</sub> SCH <sub>2</sub>	Na	55	SMTZ		Na
2	SMTZ	CNCH <sub>2</sub> SOCH <sub>2</sub>	Na				
3	SMTZ	CNCH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
4	OAc	CF <sub>3</sub> CH <sub>2</sub> SCH <sub>2</sub>	Na				
5	OAc	CF <sub>3</sub> CH <sub>2</sub> SOCH <sub>2</sub>	Na	56	SMTZ		Na
6	OAc	CF <sub>3</sub> CH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
7	SMTD	CF <sub>3</sub> CH <sub>2</sub> SCH <sub>2</sub>	Na				
8	SMTD	CF <sub>3</sub> CH <sub>2</sub> SOCH <sub>2</sub>	Na				
9	SMTD	CF <sub>3</sub> CH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na	57	SMTZ		Na
0	SMTZ	CF <sub>3</sub> CH <sub>2</sub> SCH <sub>2</sub>	Na				
1	SMTZ	CF <sub>3</sub> CH <sub>2</sub> SOCH <sub>2</sub>	Na				
2	SMTZ	CF <sub>3</sub> CH <sub>2</sub> SO <sub>2</sub> CH <sub>2</sub>	Na				
3	OAc	NaO <sub>2</sub> C(CH <sub>2</sub> ) <sub>3</sub> CONH(Ph)	Na				
4	OAc	HO <sub>2</sub> CCH <sub>2</sub> SCH <sub>2</sub> CONH(Ph)	H	58	SMTZ		Na
5	OAc	NaO <sub>2</sub> CCH <sub>2</sub> OCH <sub>2</sub> CONH(Ph)	Na				

TABLE 2 (continued)

No.	X	Y	M	No.	X	Y
59	SMTZ		Na			
a	SMTD			S-4MST		
				S-5MST		
	SHLT			SMTZ		
				SHTZ		
						

$$W_{ij} = \left[ \sum_{k=1}^m (x_{ik} - x_{jk})^2 \right]^{1/2}$$

where  $x_{ik}$  is the logarithm of  $1/\text{MIC}$  of the  $i$ -th sample against the  $k$ -th bacterium. If only maximum concentration values of the compounds tested without inhibition are reported, the values are doubled for this analysis. The results are presented in Table 3 and Fig. 2. Table 3 is the computer output for the MST in the 4-dimensional space based on antibacterial activity against four gram-negative bacteria, which contains the node (sample) pairs and the corresponding edge weights.

Figure 2 illustrates a two-dimensional map of cephalosporins in the MST. This map was obtained by using a linear mapping method (Karhunen—Loève transform) which has already been described in detail [4]. The weighted MST shows the valid relation of the neighbours in the original data space. Thus, from the weighted MST shown in Table 3 and Fig. 2, more detailed information can be obtained from the data.

The use of a "main diameter" on the MST has been suggested by Zahn

TABLE 3

Computer output for node pairs and edge weights in the MST of 62 cephalosporins

USED MEASUREMENT FOLLOWS					
E.C.	K.P.	SAL.	E.A.		
MINIMAL SPANNING TREE ANALYSIS CONDITION					
SAMPLE NUMBER	:	62			
SAMPLE DIMENSION	:	4			
END CLASS	:	1			
NODE PAIRS AND EDGE WEIGHTS IN MINIMAL SPANNING TREE					
(48 51)	.34440E-01	(24 30)	.49098E-01	(30 39)	.54861E-01
(47 53)	.85060E-01	(44 48)	.10364	(37 17)	.12656
(50 59)	.19820	(42 34)	.20458	(37 48)	.25823
(24 37)	.26102	(54 35)	.26151	(30 58)	.26202
(26 44)	.26674	(44 49)	.26697	(23 56)	.27132
(53 62)	.27795	(61 16)	.29029	(42 60)	.30822
(46 56)	.31031	(45 56)	.31439	(38 16)	.32126
(52 54)	.32210	(55 62)	.32342	(46 58)	.32474
(38 54)	.32985	(32 3)	.34493	(25 2)	.35315
(49 59)	.35894	(52 57)	.38438	(26 36)	.38591
(25 5)	.38633	(29 38)	.38802	(33 38)	.38812
( 9 7)	.39862	(38 40)	.40019	(14 11)	.40627
(32 40)	.41363	( 9 8)	.42477	(34 17)	.42578
(27 10)	.42602	(26 10)	.42606	(13 12)	.42632
(33 21)	.43078	( 6 8)	.43260	(31 15)	.43285
( 2 7)	.43563	( 2 4)	.44073	(54 56)	.44908
(32 31)	.45400	(43 18)	.46785	(19 18)	.47031
(60 12)	.47540	(47 57)	.48679	(11 12)	.50148
(41 47)	.50531	(13 22)	.51134	( 1 19)	.53317
( 1 7)	.54298	( 5 11)	.56267	(49 20)	.65318
(25 28)	.67352				

[5]. This main diameter is the path with the largest number of edges or nodes. The main diameter of the MST in Fig. 2 is shown in Fig. 3; this diameter contains 27 nodes. It may be assumed that these nodes are representative nodes of the whole data set. The effect of substituents at the *x*-position on the antibacterial activity was then examined. The substituents are acetoxy (OAc), methylthiadiazolethio (SMTD), methyltetrazolethio (SMTZ) and other substituents (X) (Table 2). Figure 4 shows the types of substituents at the *x*-position of the compounds on the main diameter. The symbols A, D, Z, and X refer to cephalosporins in which the substituents at the *x*-position are, respectively, OAc, SMTD, SMTZ, and other substituents.

The diameter was divided into three sets, termed set 1, set 2 or set 3. These sets were divided in a manner such that the number of nodes in each

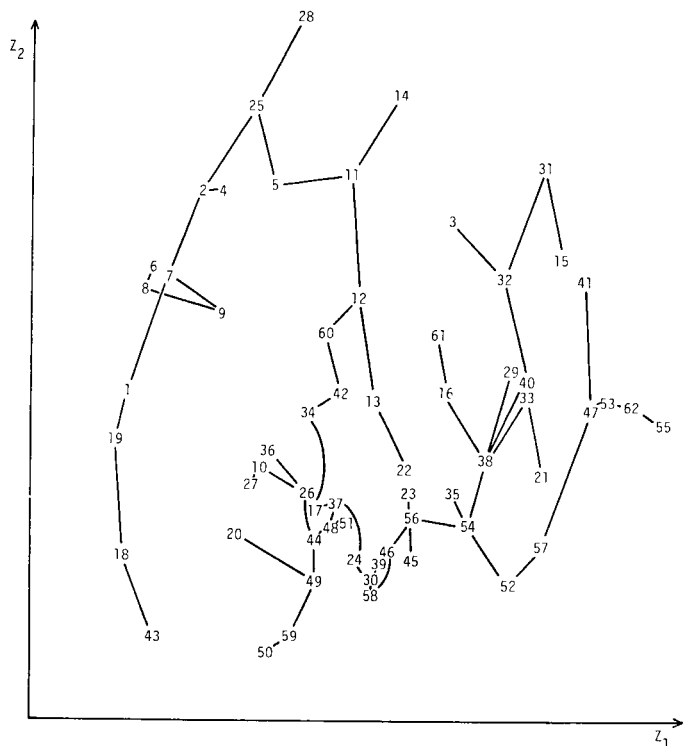


Fig. 2. MST for the antibacterial spectra data of 62 cephalosporins.

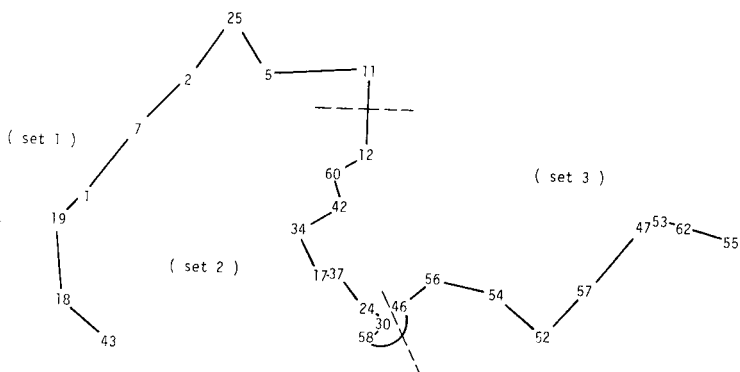


Fig. 3. Main diameter on the MST of 62 cephalosporin derivatives.

[ A - A - A - A - X - D - A - X - D ]	.....	set 1
[ - D - A - Z - A - A - D - Z - D - Z ]	.....	set 2
[ - A - Z - Z - Z - Z - Z - A - Z - Z - Z ]	.....	set 3

Fig. 4. Main diameter with the substituent at the x-position. A, acetoxy; D, SMTD; Z SMTZ; X, other substituents.



set was as equal as possible. In set 1, as shown in Fig. 4, acetyl predominates with respect to occurrence of the three substituents; in set 2, OAc, SMTD, SMTZ occur equally; and SMTZ predominates in set 3. In contrast, the mean activity value is lowest (5.617) for set 1, increasing to 7.207 for set 2 and 9.497 for set 3. It appears that the results show a good fit to the results reported previously [4] with the same data by using a hierarchic cluster analysis and the Free-Wilson method. With the present method, the results suggest that the main diameter provides a representative data structure for the whole data set.

## BENZODIAZEPINE DERIVATIVES

The MST approach, in combination with a nonlinear mapping (NLM) method, was applied to an activity study of benzodiazepine derivatives (II) which has been discussed in detail elsewhere [11, 12]. Physicochemical parameters [13] and dipole moments [11] were used for constructing the MST of the compounds; the physicochemical parameters were the Hansch  $\pi$  value, the Hammett index  $\sigma_m$ , and the molar refractivity MR (Table 4). In order to weight all parameters equally, the parameters were autoscaled to give zero mean and unit standard deviation. The MST consisting of 38 compounds was then sought in the 8-dimensional data space. The explained variances of the K-L plot in two- and three-spaces [2] are only 54% and 67% of the total variance, respectively. Therefore, the K-L plot was not applicable here for a visual display and the NLM method was employed. The MST is shown in the nonlinear map (Fig. 5).

As shown in Fig. 5, inactive compounds (7, 8, 12, 35, 37) appeared separately in the different areas. It could be suggested that the structure of the data set is asymmetric [14]. It seems that several factors are involved in the inactivity. Further, the active compounds were separated into two groups, more active class ( $-\log ED_{50} > 1$ ) and less active class ( $-\log ED_{50} < 1$ ), and the positions or area where those compounds are located were examined. It appears from Fig. 5 that the more active compounds have similar physicochemical properties. The most significant parameters will be obtained by regression analysis. In such cases, the MST may be useful because it can describe information in the parameter space.

## DISCUSSION

This investigation of the relationships among the points in the multi-dimensional data spaces of cephalosporins and benzodiazepines has shown that the mapping method is suitable for displaying multidimensional data structures and provides some information directly about the feature space. However, if the intrinsic dimensionality of the data is not low, it is very difficult to visualize the data structure by using only mapping methods. In such a case, the MST technique in combination with mapping methods is

TABLE 4

## Benzodiazepine derivatives

No.	Substituents			Physicochemical parameters								log (1/ED <sub>50</sub> )
	X	Y	Z	$\pi_x$	$\pi_y$	$\sigma_{my}$	MR <sub>y</sub>	$\pi_z$	$\sigma_{mz}$	MR <sub>z</sub>	$\mu$	
1	H	H	H	0.00	0.00	0.00	1.03	0.00	0.00	1.03	5.13	0.19
2	H	F	H	0.00	0.14	0.34	0.92	0.00	0.00	1.03	3.64	0.22
3	H	Cl	H	0.00	0.71	0.37	6.03	0.00	0.00	1.03	3.02	0.55
4	H	CN	H	0.00	-0.57	0.56	6.33	0.00	0.00	1.03	2.80	0.54
5	H	NO <sub>2</sub>	H	0.00	-0.28	0.71	7.36	0.00	0.00	1.03	1.75	1.27
6	H	CF <sub>3</sub>	H	0.00	0.88	0.43	5.02	0.00	0.00	1.03	2.64	1.48
7	H	CH <sub>3</sub>	H	0.00	0.56	-0.07	5.65	0.00	0.00	1.03	5.28	Inact.
8	H	N(CH <sub>3</sub> ) <sub>2</sub>	H	0.00	0.18	-0.15	15.55	0.00	0.00	1.03	5.04	Inact.
9	H	SCH <sub>3</sub>	H	0.00	0.61	0.15	13.82	0.00	0.00	1.03	4.78	0.20
10	H	SC <sub>2</sub> H <sub>5</sub>	H	0.00	1.07	0.18	18.42	0.00	0.00	1.03	4.94	-
11	H	SOCH <sub>3</sub>	H	0.00	-1.58	0.52	13.70	0.00	0.00	1.03	3.45	0.23
12	H	SO <sub>2</sub> CH <sub>3</sub>	H	0.00	-1.63	0.60	13.49	0.00	0.00	1.03	6.06	Inact.
13	CH <sub>3</sub>	Cl	H	0.56	0.71	0.37	6.03	0.00	0.00	1.03	2.85	0.97
14	CH <sub>3</sub>	NO <sub>2</sub>	H	0.56	-0.28	0.71	7.36	0.00	0.00	1.03	1.80	1.46
15	CH <sub>3</sub>	CN	H	0.56	-0.57	0.56	6.33	0.00	0.00	1.03	2.21	1.13
16	CH <sub>3</sub>	SCH <sub>3</sub>	H	0.56	0.61	0.15	13.82	0.00	0.00	1.03	4.70	0.77
17	CH <sub>3</sub>	N(CH <sub>3</sub> ) <sub>2</sub>	H	0.56	0.18	-0.15	15.55	0.00	0.00	1.03	4.96	0.29
18	CH <sub>3</sub>	Cl	F	0.56	0.71	0.37	6.03	0.14	0.34	0.92	2.09	1.48
19	CH <sub>3</sub>	Cl	Cl	0.56	0.71	0.37	6.03	0.71	0.37	6.03	1.89	0.90
20	CH <sub>3</sub>	NO <sub>2</sub>	F	0.56	-0.28	0.71	7.36	0.14	0.34	0.92	0.87	2.49
21	CH <sub>3</sub>	NO <sub>2</sub>	Cl	0.56	-0.28	0.71	7.36	0.71	0.37	6.03	1.29	1.04
22	CH <sub>3</sub>	NO <sub>2</sub>	CF <sub>3</sub>	0.56	-0.28	0.71	7.36	0.88	0.43	5.02	2.17	1.86
23	CH <sub>3</sub>	H	F	0.56	0.00	0.00	1.03	0.14	0.34	0.92	4.53	0.12
24	CH <sub>3</sub>	F	F	0.56	0.14	0.34	0.92	0.14	0.34	0.92	2.92	-0.14
25	CH <sub>3</sub>	N(CH <sub>3</sub> ) <sub>2</sub>	Cl	0.56	0.18	-0.15	15.55	0.71	0.37	6.03	4.27	0.33
26	H	Br	H	0.00	0.86	0.39	8.88	0.00	0.00	1.03	1.75	1.10
27	H	Br	F	0.00	0.86	0.39	8.88	0.14	0.34	0.92	0.34	1.52
28	H	Cl	F	0.00	0.71	0.37	6.03	0.14	0.34	0.92	2.30	0.85
29	H	Cl	Br	0.00	0.71	0.37	6.03	0.86	0.39	8.88	2.44	0.84
30	H	Cl	Cl	0.00	0.71	0.37	6.03	0.71	0.37	6.03	2.10	0.48
31	H	Cl	OCH <sub>3</sub>	0.00	0.71	0.37	6.03	-0.02	0.12	7.87	2.65	0.00
32	H	Cl	CH <sub>3</sub>	0.00	0.71	0.37	6.03	0.56	-0.07	5.65	3.10	0.27
33	H	CN	F	0.00	-0.57	0.56	6.33	0.14	0.34	0.92	1.24	0.57
34	H	NO <sub>2</sub>	F	0.00	-0.28	0.71	7.36	0.14	0.34	0.92	0.75	1.87
35	H	NO <sub>2</sub>	Cl	0.00	-0.28	0.71	7.36	0.71	0.37	6.03	1.18	Inact.
36	H	NO <sub>2</sub>	CF <sub>3</sub>	0.00	-0.28	0.71	7.36	0.88	0.43	5.02	2.13	0.54
37	H	NO <sub>2</sub>	NO <sub>2</sub>	0.00	-0.28	0.71	7.36	-0.28	0.71	7.36	3.81	Inact.
38	H	CF <sub>3</sub>	CF <sub>3</sub>	0.00	0.88	0.43	5.02	0.88	0.43	5.02	2.78	1.09

very useful in understanding the multidimensional feature space. The main diameter of the MST can provide a representative subset of the whole data. Then, when cluster analysis is applied, it may be possible to obtain explanatory information for a characteristic cluster.

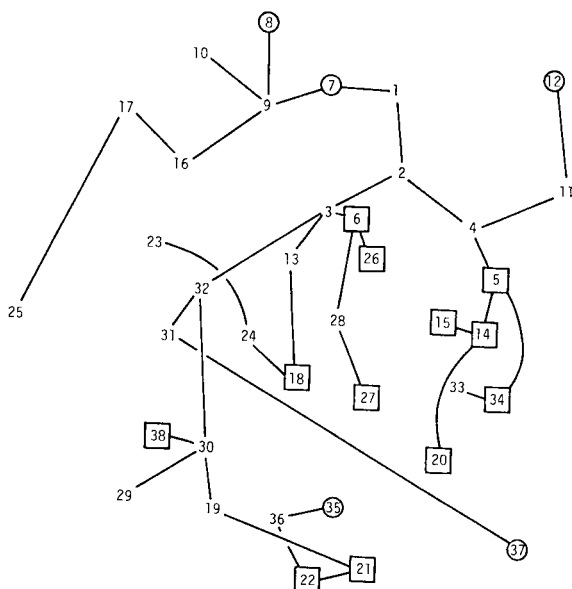


Fig. 5. MST for the physicochemical properties of 38 benzodiazepines: (○) inactive; (□) more active.

## REFERENCES

- 1 B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, 94 (1972) 5632.
- 2 B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, 95 (1973) 686.
- 3 Y. Takahashi, Y. Miyashita, H. Abe, S. Sasaki, Y. Yotsui and M. Sano, *Anal. Chim. Acta*, 122 (1980) 241.
- 4 Y. Miyashita, Y. Takahashi, Y. Yotsui, H. Abe and S. Sasaki, *CODATA Bull.*, 41 (1981) 37.
- 5 C. T. Zahn, *IEEE Trans. Comput.*, C-20 (1971) 68.
- 6 J. B. Kruskal, Jr., *Proc. Am. Math. Soc.*, 7 (1956) 48.
- 7 R. M. DeMarinis, J. R. E. Hoover, L. L. Lam, J. V. Uri, J. R. Guarini, L. Phillips, P. Actor and J. A. Weisbach, *J. Med. Chem.*, 19 (1976) 754.
- 8 R. M. DeMarinis, J. C. Boehm, G. L. Dunn, J. R. E. Hoover, J. V. Uri, J. R. Guarini, L. Phillips, P. Actor and J. A. Weisbach, *J. Med. Chem.*, 20 (1977) 30.
- 9 R. M. DeMarinis, J. C. Boehm, J. V. Uri, J. R. Guarini, L. Phillips and G. L. Dunn, *J. Med. Chem.*, 20 (1977) 1164.
- 10 L. H. Sternbach, L. O. Randall, R. Banziger and H. Lehr in A. Berger (Ed.), *Drugs Affecting the Central Nervous System*, Vol. 2, M. Dekker, New York, 1968, p. 237.
- 11 T. Blair and G. A. Webb, *J. Med. Chem.*, 20 (1977) 1206.
- 12 I. Lukovits and A. Lopata, *J. Med. Chem.*, 23 (1980) 449.
- 13 C. Hansch and A. J. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*, J. Wiley, New York, 1980.
- 14 W. J. Dunn III and S. Wold, *J. Med. Chem.*, 23 (1980) 595.

## PATTERN RECOGNITION FOR THE STUDY OF STRUCTURE—ACTIVITY RELATIONSHIPS

### Uses of the Adaptive Least-squares Method and Linear Discriminant Analysis

IKUO MORIGUCHI\*, KATSUICHIRO KOMATSU and YASUO MATSUSHITA

*School of Pharmaceutical Sciences, Kitasato University, Shirokane 5-chome,  
Minato-ku, Tokyo 108 (Japan)*

(Received 23rd January 1981)

#### SUMMARY

The adaptive least-squares method (ALS) and linear discriminant analysis (LDA) were applied to structure—activity correlation studies including the antitumor activity of mitomycin derivatives, the relative binding affinities of 100 steroids for five receptors, assignment of the pharmacological category of 80 diarylmethane-derived drugs, and discrimination of the adverse reaction of 98 miscellaneous drugs that may induce liver and/or blood diseases. Generally, more satisfactory results were obtained by the use of ALS than by LDA, both in recognition and in leave-one-out predictions. However, LDA was not always inferior to ALS in the applications, especially those related to classification of independent categories.

In spite of recent developments in the field of life sciences, theoretical prediction of the biological activity of a compound from its molecular structure still remains impossible in most cases, and empirical studies of structure—activity relationships play an important role in current computer-assisted drug design.

In the empirical studies of structure—activity relations, molecular structures of drugs are usually quantified by physico-chemical parameters such as hydrophobic, electronic, and steric constants, and by substructural descriptors, such as structural fragments and functional groups. Biological activity is expressed by the level or the kind of action. When the potency level is observed in an interval scale, multiple-regression analysis such as the Hansch approach [1, 2] is preferred for the study, and the structure—activity relationship is formulated as a regression equation. However, when the potency level is recorded in an ordinal (sequential) scale, or when the activity is given by the kind of action, techniques belonging to pattern classification [3–6] may be applicable. In this case, the relationship between structure and activity is usually expressed by discriminant functions.

Pattern classification includes various techniques. For the purpose of structure—activity correlations, however, the method should be supervised learning, and the structure—activity relationship should be formulated

mathematically, e.g., as discriminant functions. Further, the method should be effective even in cases where substances are not completely separable. From this point of view, two methods, the adaptive least-squares method (ALS) [7, 8] and linear discriminant analysis (LDA) [9], were chosen for structure-activity correlations in this work. As is well known, LDA is constructed on a parametric basis for classifying substances into independent categories, whereas ALS is nonparametrically designed for classification of ordered categories, or graded classes.

## METHODS

### Adaptive least-squares method

The adaptive least-squares method (ALS) makes decisions for multicategory classification by a single discriminant function which is formulated by a feedback adaptation procedure in a simple linear form:

$$L = w_0 + w_1x_1 + w_2x_2 + \dots + w_px_p \quad (1)$$

where  $L$  is the discriminant score for classification,  $x_k$  ( $k = 1, 2, \dots, p$ ) is the  $k$ th descriptor for the structure, and  $w_k$  ( $k = 0, 1, \dots, p$ ) is the weight coefficient. The value of  $w_k$  is determined by least-squares calculation using the forcing factor as illustrated in Fig. 1. The forcing factor starts with a starting score, and is iteratively adapted by using a correction term.

The starting score,  $a_j$  ( $j = 1, 2, \dots, m$  in the  $m$ -group case), was assumed in the same manner as previously described [8] using the ridity [10], as

$$a_j = 4 \text{ ridit}(j) - 2 = 4 \left( \sum_{g=1}^{j-1} n_g + n_j/2 \right) / n - 2 \quad (2)$$

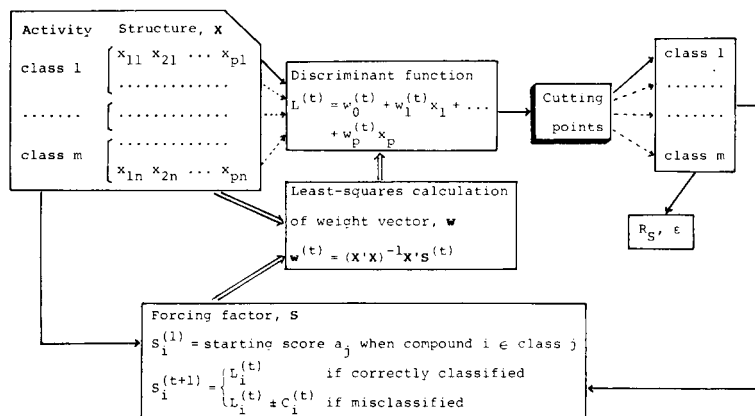


Fig. 1. Outline of ALS.

where  $n$  is the total number of compounds, and  $n_g$  and  $n_j$  are the size of groups  $g$  and  $j$ , respectively.

The cutting point,  $b_j$  ( $j = 1, 2, \dots, m - 1$ ), between classes was fixed to be the midpoint between the starting scores.

The correction term,  $C_i^{(t)}$ , for substance  $i$  at the  $t$ th iteration was given by

$$C_i^{(t)} = 0.1/[\alpha + \delta_i^{(t)}]^2 + \beta[L_i^{(t)} - S_i^{(1)}]^2 \quad (3)$$

where  $L_i^{(t)}$  is the discriminant score for substance  $i$  at the  $t$ th iteration,  $S_i^{(1)}$  is the initial forcing factor equal to the starting score for substance  $i$ , and  $\delta_i^{(t)}$  is the deviation defined by

$$\delta_i^{(t)} = |L_i^{(t)} - b_k| \quad (4)$$

In eqn. (4),  $b_k$  is the cutting point (nearer to  $L_i^{(t)}$ ) for the observed class for substance  $i$ .

The constants  $\alpha$  and  $\beta$  in eqn. (3) were empirically taken to be  $\alpha = 0.45$  and  $\beta = 0.01, 0.03$ , and  $0.05$ . In each run using these values, the adaptive iteration was performed 20 times, and the best discriminant function was selected.

As the criteria of the best discrimination, the Spearman rank correlation coefficient with a correction of many ties [11],  $R_s$ , and the  $\epsilon$  value [8] corresponding to the mean square of empirical errors were used.

#### *Linear discriminant analysis*

The Rao method [9, 12] generally used for several groups was employed. Prior probabilities were taken to be proportional to the group size. Because equality of the within-group covariance matrices was not fulfilled in most data in this study, statistical tests for the significance of classification were not done.

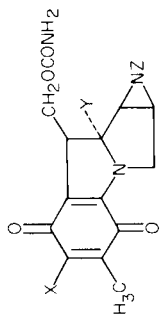
#### *Selection of descriptors*

As a measure of the contribution of descriptors to the discriminant score in ALS calculations, a contribution index, which is the product of the weight coefficient and the standard deviation, was used for each descriptor. For the preliminary selection of descriptors, a backward stepwise elimination was carried out on the basis of the contribution index at the first iteration for ALS, and the generalized Mahalanobis'  $D^2$  [12] for LDA. Because these criteria for the selection were only tentative, the results were used for a rough grading of the descriptors into the following three ranks: (a) key descriptors to be included; (b) descriptors ( $\leq 10$ ) to be further selected using their every-possible-combination added to the key descriptors; and (c) unwanted descriptors to be removed. The final set of descriptors was selected on the basis of the result of the leave-one-out prediction.

*Computation.* All computations were done with an OKITAC system 50 model 40 computer using double precision. All of the programs used were written in JIS 7000 FORTRAN.

TABLE 1

Structural features and activity of mitomycin derivatives



	X	Y	Z	$\sigma_m-X^a$	$V_{W-X}$	$\sigma^*Y^b$	$B_{1-Z}^c$	$B_{4-Z}^c$	Solid sarcoma <sup>d</sup>		Ascites sarcoma <sup>e</sup>	
									Obs. <sup>f</sup>	Calc. <sup>g</sup>	Pred. <sup>h</sup>	Obs. <sup>f</sup>
1 <sup>i</sup>	NH <sub>2</sub>	OMe	H	-0.16	0.177	1.81	1.00	1.00	3	3	3	3
2	NHEt	OMe	H	-0.24	0.493	1.81	1.00	1.00	3	3	3	2
3 <sup>j</sup>	NH <sub>2</sub>	OMe	Me	-0.16	0.177	1.81	1.52	2.04	3	3	3	3
4	NH <sub>2</sub>	OMe	Et	-0.16	0.177	1.81	1.52	2.97	3	3	2	2
5	NH <sub>2</sub>	OMe	Ac	-0.16	0.177	1.81	1.90	2.93	3	3	2	2
6	NH <sub>2</sub>	OH	Me	-0.16	0.177	1.55	1.52	2.04	3	3	2	2
7	NMe <sub>2</sub>	OMe	H	-0.15	0.441	1.81	1.00	1.00	3	3	3	3
8	NH <sub>2</sub>	OMe	COPh-o-Cl	-0.16	0.177	1.81	2.36	5.98	2	2		
9	NH <sub>2</sub>	OMe	COPh-p-Cl	-0.16	0.177	1.81	2.36	5.98	2	2		
10	NHPh	OMe	H	-0.12	0.892	1.81	1.00	1.00	2	2	2	2
11 <sup>k</sup>	OMe	OMe	H	0.12	0.304	1.81	1.00	1.00	2	2	3	3
12	OMe	OMe	Me	0.12	0.304	1.81	1.52	2.04	2	2	2	3
13 <sup>l</sup>	OMe	OH	Me	0.12	0.304	1.55	1.52	2.04	1	1	2	2
14	NH <sub>2</sub>	H	Me	-0.16	0.177	0.49	1.52	2.04	1	1	1	1
15	NH <sub>2</sub>	OMe	SO <sub>2</sub> Me	-0.16	0.777	1.81	2.11	3.15	1	2	3	2
16	OMe	H	Me	0.12	0.304	0.49	1.52	2.04	1	1	1	1

<sup>a</sup>From ref. 14. <sup>b</sup>From ref. 16. <sup>c</sup>From ref. 17. <sup>d</sup>Sarcoma 180. <sup>e</sup>Hirosaki ascites sarcoma. <sup>f</sup>From ref. 13. <sup>g</sup>ALS recognition. <sup>h</sup>ALS leave-one-out prediction. <sup>i</sup>Mitomycin C. <sup>j</sup>Porfiromycin. <sup>k</sup>Mitomycin A. <sup>l</sup>Mitomycin B.

## RESULTS AND DISCUSSION

The ALS and LDA methods were applied to four structure—activity correlation studies. The first two were on the problems of ordered-categorical classification, and the other two studies were on the independent-categorical classification.

*Antitumor activity of mitomycin derivatives*

The ALS and LDA methods were applied to the structure—activity correlation of mitomycin derivatives [13]. Table 1 shows the structural features and antitumor activity of the mitomycin derivatives. The five descriptors proved to be effective. They were the Hammett constant for the *meta*-position,  $\sigma_m$  [14] and the van der Waals volume,  $V_w$  [15] for the substituent X, the polar effect,  $\sigma^*$  [16] for Y, and the Sterimol parameters  $B_1$  and  $B_4$  [17] for Z. Their values are listed in Table 1. Activity ratings for 16 compounds were observed against solid sarcoma in mice, and 14 compounds against ascites sarcoma.

The three-group discrimination was done by using ALS and LDA. The results of ALS recognition and leave-one-out prediction for individual compounds are also shown in Table 1. Table 2 lists the discriminant functions and Spearman's constants for ALS and LDA classifications.

The structure—activity relationship was clearly quantified in the ALS discriminant functions. Electron-attracting and bulky X-substituents reduce the activity, polar Y-substituents enhance the strength of action, and large Z-substituents decrease the potency. These results may support the mechanism of action proposed by Murakami [18], who suggested a crosslinkage formation with the coupled bases in DNA.

TABLE 2

ALS and LDA discriminant functions with mitomycin derivatives

Descriptor	Solid sarcoma 180 in mice				Ascites Hirotsaki sarcoma in mice			
	ALS	LDA			ALS	LDA		
		(1) <sup>a</sup>	(2) <sup>b</sup>	(3) <sup>c</sup>		(1) <sup>a</sup>	(2) <sup>b</sup>	(3) <sup>c</sup>
$\sigma_m-X$	-4.329	40.35	40.23	14.90				
$V_w-X$	-2.568	44.34	47.74	30.20	-2.110	23.98	24.14	16.13
$\sigma^*_Y$	1.563	4.27	10.50	11.89	1.744	8.51	21.01	22.73
$B_1-Z$	-1.430	23.48	23.49	14.46				
$B_4-Z$					-0.546	7.86	8.08	5.45
Constant	-0.027	-28.21	-38.08	-23.79	-1.081	-16.29	-32.53	-27.97
$R_S^d$ {								
recogn.	0.969	0.946			0.878	0.795		
predict. <sup>e</sup>	0.833	0.601			0.706	0.667		

<sup>a</sup>For poorly active compounds. <sup>b</sup>For moderately active compounds. <sup>c</sup>For very active compounds. <sup>d</sup>Spearman's rank correlation coefficient. <sup>e</sup>Using leave-one-out technique.



In contrast, LDA produced three discriminant functions corresponding to three activity classes against each sarcoma, and the structure—activity relations comprised were somewhat indistinct.

The Spearman's rank correlation coefficients (Table 2) indicate that ALS provides more satisfactory results than LDA, both in the recognition and the leave-one-out prediction.

#### *Relative binding affinities of steroids for five receptors*

The relative binding affinities of steroids for five receptors, i.e., estrogen (ES), progestin (PG), androgen (AND), mineralocorticoid (MIN), and glucocorticoid (GLU) receptors, were described in an ordinal scale by Raynaud et al. [19]. The steroid compounds used in this study were 24 estradiol derivatives, 26 progesterone derivatives, 26 testosterone derivatives, and 24 mineralo- and gluco-corticoids.

Table 3 shows several examples of the steroids along with their affinity ratings which were somewhat simplified for this study as described in the footnote. From Table 3, it can be understood that some steroids bind to several receptor populations, so such steroids can cause highly undesirable side-effects in medical therapy.

The ALS discriminant functions for receptor bindings are listed in Table 4. The descriptors used were dummy variables for double bonds, various substituents, and phenolic rings, the polar effects for substituents at the 9- and the 11-positions, and the number of the carbon atoms in the 13- $\beta$ -alkyl groups. In the estrogen receptor binding, only three descriptors gave a high rate of classification. The Spearman's rank correlation coefficient was excellent (0.996). This suggests that the structural requirement of the estrogen receptor is highly specific. However, the structural requirements of four other receptors are complicated and partly overlapping. This must be fully considered in the design of new steroid drugs in order to avoid undesirable adverse reactions.

TABLE 3

Relative binding affinities<sup>a</sup> of several steroids for five receptors

Compound	ES	PG	AND	MIN	GLU
Ethynylestradiol	+++ <sup>b</sup>	++ <sup>b</sup>	+ <sup>b</sup>	— <sup>b</sup>	+ <sup>b</sup>
19-Nor-pregna-4,9,11-triene-3,20-dione	—	+++	++	+	+
5- $\alpha$ -Dihydrotestosterone	—	—	+++	—	—
7 $\alpha$ ,17 $\alpha$ -Dimethyl-13 $\beta$ -ethyl-17-hydroxy-gona-4,9,11-trien-3-one	—	+++	+++	+++	+++
Fludrocortisone	—	—	—	+++	+++
Triamcinolone acetonide	—	+	—	+	+++

<sup>a</sup>From ref. 19. <sup>b</sup>—, <10; +, 10—50; ++, 50—125; +++, >125.

TABLE 4

ALS discriminant functions for receptor bindings

Descriptor	ES	PG	AND	MIN	GLU
Dm <sup>a</sup> for 4/5(10)-ene				0.59	0.51
Dm for 6-ene		-0.49	-0.58		0.27
Dm for 9(10)-ene		-0.60			1.26
Dm for 11-ene		0.84	0.79	0.92	
Dm for 2 or 4-Me		-1.05	-0.83		-0.35
Dm for 6-Me, F, or Cl			1.32		
Dm for 7 $\alpha$ -Me	0.35				
$\Delta\sigma^{*b}$ for 9-subst.		0.28	0.29	0.30	0.38
Dm for 11 $\beta$ -OH		-0.80	-1.00	0.72	0.84
$\Delta\sigma^*$ for 11-subst.	-0.24	-0.16	-0.43	-0.50	0.64
Dm for 16-Me		0.58		-0.96	
Dm for 16-OH or OR			-0.68		
Dm for 17 $\beta$ -COMe/COEt			-0.92	0.09	-0.88
Dm for 17 $\alpha$ -alkyl		0.26			0.78
Dm for 17 $\alpha$ -OAc		1.61		-0.79	0.93
Dm for 17-[A] <sup>c</sup> or [B] <sup>d</sup>				1.07	
Dm for 17-[C] <sup>e</sup>			-0.86	0.36	
Dm for 17 $\alpha$ -C $\equiv$ CH			-0.82		0.91
Dm for 17 $\beta$ -COCH <sub>2</sub> OR			-0.97	1.32	
Dm for 17 $\beta$ -OH		-1.07	0.63		-0.46
Dm for 17 $\alpha$ -OH		-1.39			
Dm for 17 $\alpha$ -OR				-1.76	
Dm for 17-CH <sub>2</sub> OH		-1.49			
No of C in 13 $\beta$ -alkyl					0.59
Dm for 10 $\beta$ -Me		-1.54			
Dm for phenolic A ring	2.00	-1.53	-1.44		0.46
Dm for phenolic D ring		-1.05			
Constant	-0.41	1.66	0.76	-0.83	-1.82
$R_S$ (Spearman's coeff.)	0.996	0.961	0.935	0.865	0.914

<sup>a</sup>Dummy variable. <sup>b</sup>Difference in Taft's polar constant between the substituent and hydrogen. <sup>c</sup>

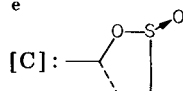
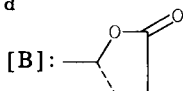
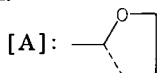


Table 5 shows the results of ALS and LDA for the receptor bindings of a hundred bioactive steroids. In all cases, ALS provides more satisfactory results than LDA, both in the recognition and the leave-one-out prediction.

#### *Discrimination of the pharmacological categories of diarylmethane-derived drugs*

Many drugs acting as anticholinergic agents, antidepressants, and antihistamines have diarylmethane-derived structures. This study attempts to discriminate the pharmacological categories according to their structures.

TABLE 5

Results of ALS and LDA with 100 bioactive steroids

Receptor	ALS					LDA				
	N <sup>a</sup>	Recogn.		Predict. <sup>b</sup>		N <sup>a</sup>	Recogn.		Predict. <sup>b</sup>	
		n <sub>mis</sub> <sup>c</sup>	R <sub>S</sub> <sup>d</sup>	n <sub>mis</sub>	R <sub>S</sub>		n <sub>mis</sub>	R <sub>S</sub>	n <sub>mis</sub>	R <sub>S</sub>
ES	3	4(0) <sup>e</sup>	0.996	6(0) <sup>e</sup>	0.993	3	6(2) <sup>e</sup>	0.939	8(2) <sup>e</sup>	0.937
PG	16	9(1)	0.961	19(1)	0.921	14	13(1)	0.942	24(2)	0.901
AND	14	14(0)	0.935	22(0)	0.900	14	20(0)	0.907	25(0)	0.887
MIN	12	18(2)	0.865	20(2)	0.850	12	19(3)	0.841	28(5)	0.765
GLU	14	18(0)	0.914	25(1)	0.861	15	22(0)	0.887	32(1)	0.822

<sup>a</sup>Number of descriptors. <sup>b</sup>Using leave-one-out technique. <sup>c</sup>Number of compounds misclassified. <sup>d</sup>Spearman's rank correlation coefficient. <sup>e</sup>The figures in parentheses are the number of compounds misclassified by more than one grade.

Table 6 lists the 80 drugs used in the study, including 35 anticholinergic agents (C), 16 antidepressants (D), and 29 antihistamines (H), cited from the literature [20, 21].

Table 7 shows some of the ALS discriminant functions. Eleven descriptors were effective for the discriminations. They were the hydrophilic effect,  $V_H$  [15], the van der Waals volume,  $V_W$  for the  $R_1$ -substituents at the 2- and

TABLE 6

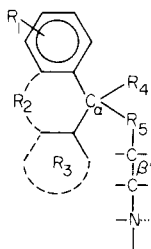
Diarylmethane-derived anticholinergic agents, antidepressants, and antihistamines used in the study

Anticholinergic agents (35 drugs)			
Adiphenine	Ambutonium	Aminopentamide	Benapryzine
Bentipimine	Benzilonium	Benzomethamine Cl	Benzopyronium Br
Benztropine	Bufenadrine	Butinoline	Buzepide methiodide
Carbatrine	Chlorphenoxamine	Clidinium Br	Clefenetamine
Diphebanil mesylate	Ethylbenztropine	Heteronium Bromide	Isopropamide I
Meclozamine	Mepenzolate Br	Methantheline Br	Methixene
Orphenadrine	Parapenzolate Br	Pipenzolate Br	Piperidolate
Piperilate	Pipoxolan	Poldine mesylate	Pridinol
Propantheline Br	Thiphenamil	Triclazate	
Antidepressants (16 drugs)			
Amitriptyline	Butriptyline	Diox adrol	Dothiepin
Doxepin	Hepzidine	Intriptyline	Maprotiline
Melitracen	Nortriptyline	Noxiptilin	Octriptyline
Pizotyline	Prothixene	Protriptyline	Tofenacin
Antihistamines (29 drugs)			
Azatadine	Bromodiphenhydramine	Brompheniramine	Carbinoxamine
Chlorcyclizine	Chlorpheniramine	Cinnarizine	Clemastine
Clobenztropine	Cyclizine	Cyproheptadine	Diphenhydramine
Diphenylpyraline	Dithiadene	Doxylamine	Embramine
Fenipirane	Meclizine	Medrylamine	Mephenhydramine
4-Methylphenhydramine	Perastine	Phenindamine	Pheniramine
Pirdonium Br	Pimethixene	Pyroxamine	Tolpropamine
Triprolidine			

TABLE 7

ALS discriminant functions with 80 diarylmethane-derived drugs

Descriptor	[D, H]/[C] <sup>a</sup>	[H, C]/[D] <sup>a</sup>	[C, D]/[H] <sup>a</sup>
1. $V_H$ (hydrophilic effect)	-1.103		0.757
2. $V_W$ for 2- $R_1$	2.921		-3.411
3. $V_W$ for 4- $R_1$			1.039
4. Length of $C_\alpha-R_5-C_{\beta'}$	0.212		-0.306
5. $Dm^b$ for $R_2$ including C	-0.911	1.137	-0.280
6. $Dm$ for $R_3 = \text{pyridyl}$		-0.329	
7. $Dm$ for $R_4 = \text{OH or OR}$	0.895		-1.153
8. $Dm$ for $R_4 = \text{CONH}_2$	2.157		-1.918
9. $Dm$ for $R_5 = \text{CO-S/O/N-}$	0.822		
10. $Dm$ for sec. N	-0.522	1.129	
11. $Dm$ for quat. N	2.524		-1.908
Constant	0.367	-0.341	0.607
$R_S$ (Spearman's coeff.)	0.878	0.928	0.838



<sup>a</sup>[C], Anticholinergic agents (35 drugs); [D], antidepressants (16 drugs); and [H], antihistamines (29 drugs). <sup>b</sup>Dummy variable.

the 4-positions, the length between  $C_\alpha$  and  $C_{\beta'}$  atoms along the bonds, and seven dummy variables concerning the  $R_2$  bridge,  $R_3$  ring,  $R_4$  and  $R_5$  groups, and the nitrogen in the side chain.

The discriminant functions shown in Table 7 are for the two-group discriminations. For the antidepressants, only three descriptors provided a high rate of discrimination. The structural requirement of the antidepressants seemed very selective. For the anticholinergic agents and antihistamines, the structural requirements were somewhat complicated. However, several features for each category are clearly shown. For example, the presence of

TABLE 8

Results of ALS and LDA with 80 diarylmethane-derived drugs

Discrimination	ALS				LDA					
	$N^a$	Recogn.		Predict. <sup>b</sup>		$N^a$	Recogn.		Predict. <sup>b</sup>	
		$n_{\text{mis}}^c$	$R_S^d$	$n_{\text{mis}}$	$R_S$		$n_{\text{mis}}$	$R_S$	$n_{\text{mis}}$	$R_S$
[D, H]/[C] <sup>e</sup>	9	5	0.873	8	0.803	9	8	0.809	10	0.756
[H, C]/[D]	3	2	0.928	3	0.896	3	3	0.896	5	0.810
[C, D]/[H]	8	6	0.838	7	0.810	8	9	0.759	12	0.672
[D]/[H]/[C]	8	9	—	14	—	9	10	—	15	—
[H, C]/[D]	3 → 8	7	—	9	—	3 → 8	9	—	14	—
~ [H]/[C] <sup>f</sup>										

<sup>a</sup>Number of descriptors. <sup>b</sup>Using leave-one-out technique. <sup>c</sup>Number of compounds misclassified. <sup>d</sup>Spearman's rank correlation coefficient. <sup>e</sup>[C], [D] and [H] as in Table 7. <sup>f</sup>Step-wise discrimination.

a bulky substituent at the 2-position remarkably increases the anticholinergic activity, but considerably decreases the antihistaminic activity.

Table 8 shows the results of ALS and LDA with the 80 diarylmethane-derived drugs. The upper three discriminations correspond to the dichotomy shown in Table 7. The lower two cases are a simultaneous three-group classification, and a step-wise discrimination.

In the simple and step-wise division into two contrasting classes, ALS provided more satisfactory results than LDA. In the simultaneous three-group classification, however, LDA was not inferior to ALS.

*Discrimination of the adverse reactions of miscellaneous drugs that may induce liver and/or blood diseases*

It is known [22–24] that, as undesirable side-reactions, the first 12 drugs in Table 9 may induce liver disease, the next 16 drugs may induce blood dyscrasias, and the last 70 drugs may induce both side-reactions.

The ALS discriminant functions are shown in Table 10. The 15 descriptors listed were effective for classifying the 98 drugs in the two groups. The descriptors included the van der Waals volume of the drug molecule, the numbers of the particular atoms and bonds, and six dummy variables for

TABLE 9

Miscellaneous drugs that may induce liver disease and/or blood dyscrasias<sup>a</sup> used in the study

<b>Drugs that may induce liver disease (12 drugs)</b>			
Cinchophen	Diethylstilbestrol	Ectylurea	Ethionamide
Halothane	Methandrostenolone	Methyltestosterone	Norethandrolone
Norethisterone	Norethynodrel	Triacetyloleandomycin	Vitamin K <sub>3</sub>
<b>Drugs that may induce blood dyscrasias (16 drugs)</b>			
Acetazolamide	Amphetamine	Barbital	Chlorambucil
Digitoxin	Furazolidone	Iothiouracil	Meprobamate
Nitrofurazone	Phenobarbital	Primidone	Quinacrine
Thiazosulfone	Thiouracil	Tripelennamine	Urethane
<b>Drugs that may induce both liver disease and blood dyscrasias (70 drugs)</b>			
Acetanilide	Acetohexamide	Acetylphenylhydrazine	Aminopyrine
<i>p</i> -Aminosalicylic acid	Antipyrine	Arsphenamine	Aspirin
Benzylpenicillin	Busulfan	Carbamazepine	Carbutamide
Cephalothin	Chloramphenicol	Chlorothiazide	Chlorpromazine
Chlorpropamide	Chlortetracycline	Clofibrate	Cytarabine
Dapsone	Dinitrophenol	Ethacrynic acid	Hydrochlorothiazide
Imipramine	Iproniazid	Isocarboxazid	Isoniazid
6-Mercaptopurine	Mephentoin	Mesoridazine	Metahexamide
Methimazole	Methotrexate	Methyldopa	Neosphenamine
Nitrofurantoin	Oxyphenbutazone	Pamaquine	Perphenazine
Phenacemide	Phenacetin	Phenelzine	Phenindion
Phenylbutazone	Phenylhydrazine	Phenytion	Polythiazide
Primaquine	Probenecid	Prochlorperazine	Promazine
Propylthiouracil	Pyrazinamide	Quinidine	Salicylazosulfapyridine
Streptomycin	Sulfacetamide	Sulfamethoxazole	Sulfamethoxypyridazine
Sulfanilamide	Sulfisoxazole	Sulfoxone	Sulpyrin
Tetracycline	Thioridazine	Tolbutamide	Tranlycypromine
Trimethadione			

<sup>a</sup>From refs. 22–24.

TABLE 10

ALS discriminant functions with 98 miscellaneous drugs

Descriptor	Liver disease	Blood dyscrasias
$V_w$	-1.530 (1.52) <sup>a</sup>	-0.585 (0.58) <sup>a</sup>
Number of C	0.248 (1.66)	0.129 (0.86)
Number of H		0.054 (0.55)
Number of F	0.449 (0.19)	
Number of non-ring C		-0.194 (0.54)
Number of non-ring O	0.232 (0.51)	0.265 (0.58)
Number of non-ring S	0.141 (0.08)	
Number of C—C	-0.081 (0.55)	-0.152 (1.03)
Number of C=C (aromatic)		-0.090 (0.38)
Dm <sup>b</sup> for tert. amine		0.867 (0.34)
Dm for $\text{>NCOO—}$	-1.327 (0.19)	
Dm for $\text{—CH=NNCO—}$	-0.593 (0.10)	
Dm for aliphatic $\text{>C=O}$	0.850 (0.25)	-1.072 (0.31)
Dm for $\text{—CONCONCO—}$	-1.586 (0.23)	
Dm for $\text{—SO}_3\text{H}$		-1.667 (0.24)
Constant	-0.212	-0.081
$R_S$ (Spearman's coeff.)	0.763	1.000

<sup>a</sup>The figures in parentheses are the contribution index. <sup>b</sup>Dummy variable.

several functional groups. The structural requirements were rather complicated. The ALS discriminant functions may suggest that small molecules rich in carbon and aliphatic oxygen atoms induce liver disease, and that the presence of the saturated carbon—carbon bonds reduces blood dyscrasias.

The overall results of the discriminations are shown in Table 11. Although ALS provides more satisfactory results than LDA in the recognitions, LDA is by no means inferior in the leave-one-out predictions in this case.

In conclusion, ALS can be generally expected to provide clearer relationships and more satisfactory results than LDA, especially for ordered-categorical classifications.

The details of each study will be published separately elsewhere.

TABLE 11

Results of ALS and LDA with 98 miscellaneous drugs

Category	ALS				LDA					
	$N^a$	Recogn.		Predict. <sup>b</sup>		$N^a$	Recogn.		Predict. <sup>b</sup>	
		$n_{\text{mis}}^c$	$R_S^d$	$n_{\text{mis}}$	$R_S$		$n_{\text{mis}}$	$R_S$	$n_{\text{mis}}$	$R_S$
Liver disease	10	6	0.763	10	0.574	10	10	0.574	11	0.521
Blood dyscrasias	10	0	1.000	8	0.571	10	4	0.798	8	0.554

<sup>a-d</sup>As for Table 8.

## REFERENCES

- 1 C. Hansch, R. M. Muir, T. Fujita, P. Maloney, E. Geiger and M. Streich, *J. Am. Chem. Soc.*, 85 (1963) 2817.
- 2 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86 (1964) 1616.
- 3 P. C. Jurs and T. L. Isenhour, *Chemical Application of Pattern Recognition*, Wiley, New York, 1975.
- 4 I. Moriguchi, in T. Fujita (Ed.), *Structure—Activity Relationships*, Nankodo, Tokyo, 1979, p. 281 (in Japanese).
- 5 G. L. Kirschner and B. R. Kowalski, in E. J. Ariëns (Ed.), *Drug Design*, Vol. VIII, Academic, New York, 1979, p. 73.
- 6 A. J. Stuper, W. E. Brügger and P. C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function*, Wiley, New York, 1979.
- 7 I. Moriguchi and K. Komatsu, *Chem. Pharm. Bull.*, 25 (1977) 2800, 3440 (errata).
- 8 I. Moriguchi, K. Komatsu and Y. Matsushita, *J. Med. Chem.*, 23 (1980) 20.
- 9 C. R. Rao, *Advanced Statistical Methods in Biometric Research*, Wiley, New York, 1952.
- 10 I. D. J. Bross, *Biometrics*, 14 (1958) 18.
- 11 A. L. Delaunoy (Ed.), *Biostatistics in Pharmacology*, Vol. 2, Pergamon, Oxford, 1973, p. 943.
- 12 H. Iguchi, *Multivariate Analysis and Computer Programs*, Nikkan Kogyo Shinbunsha, Tokyo, 1972, p. 103 (in Japanese).
- 13 S. Kinoshita, K. Uzu, K. Nakano, S. Shimizu, T. Takahashi and M. Matsui, *J. Med. Chem.*, 14 (1971) 103.
- 14 C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikaitani and E. J. Lien, *J. Med. Chem.*, 16 (1973) 1207.
- 15 I. Moriguchi, Y. Kanada and K. Komatsu, *Chem. Pharm. Bull.*, 24 (1976) 1799.
- 16 Y. C. Martin, *Quantitative Drug Design*, M. Dekker, New York, 1978, p. 377.
- 17 A. Verloop, W. Hoogenstraaten and J. Tipker, in E. J. Ariëns (Ed.) *Drug Design*, Vol. VII, Academic, New York, 1976, p. 165.
- 18 H. Murakami, *J. Theor. Biol.*, 10 (1966) 236.
- 19 J. P. Raynaud, T. Ojasoo, M. M. Bouton and D. Philibert, in E. J. Ariëns (Ed.), *Drug Design*, Vol. VIII, Academic, New York, 1979, p. 169.
- 20 A. F. Harms, W. Hespe, W. Th. Nauta, R. F. Rekker, H. Timmerman and J. de Vries, in E. J. Ariëns (Ed.), *Drug Design*, Vol. VI, Academic, New York, 1975, p. 1.
- 21 T. Z. Csáky, *Cutting's Handbook of Pharmacology*, 6th edn., Appleton-Century-Crofts, New York, 1979.
- 22 E. M. Martin, *Hazards of Medication*, J. B. Lippincott, Philadelphia, 1971, p. 321.
- 23 A. Wade (Ed.), *Martindale: The Extra Pharmacopoeia*, 27th edn., Pharmaceutical Press, London, 1977.
- 24 H. Ozawa, in H. Abe, K. Mashimo, K. Sambe and H. Ozawa (Ed.), *Practice in Clinical Medication*, 2nd edn., Daiichi Seiyaku, Tokyo, 1978, Part 2 (in Japanese).

## A GRAPH-THEORETIC APPROACH TO QUANTITATIVE STRUCTURE—ACTIVITY/REACTIVITY STUDIES

CHARLES L. WILKINS<sup>\*,\*\*</sup>, MILAN RANDIĆ, SHELDON M. SCHUSTER,  
RODNEY S. MARKIN, STEVEN STEINER and LONNIE DORGAN

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

(Received 23rd January 1981)

### SUMMARY

Characterization of molecular species based on the use of suitable graph invariants (graph paths, in particular) can provide a quantitative means of encoding structure; the technique is complementary to commoner approaches to studies of quantitative structure—activity relationships. Graph path encoding is here applied to quantitative studies of relationships between molecular structures and biological activity; the examples are the rates of various substrate reactions with hexokinase, and the potential opiate-like activity of enkephalin analogs.

A topic of continuing interest is the relationship between complex biomolecules, their interaction with relatively small molecules, and their functions. The complexity of these interactions is generally such that empirical approaches to their study are still essential. Several distinctive approaches to quantitative structure—activity studies have been developed. Hansch [1] is primarily responsible for the multiple linear regression approach which is similar to the somewhat related technique of Free and Wilson [2]; both are empirical schemes in which one attempts to derive parameters from measurements on known compounds with desired activity. Subsequently, the goal is to use the derived parameters for prediction of the activity of untested molecules. Pattern recognition has also been used as an empirical scheme aimed at classifying molecules on the basis of their pharmacological activity [3–6]. The empirical character of these approaches is reflected in the limited significance of the actual parameters derived. Frequently, considerably different sets of parameters will produce equally good results. Thus, the advantage of such schemes is their practicality and reasonable accuracy, rather than their ability to yield fundamental understanding of the basis of activity.

Different assumptions characterize so-called topological indices, which have also been found useful for empirically characterizing structure—activity correlations. These indices evolved from the early studies of Wiener [7] who considered isomeric variations of thermodynamic properties of alkanes.

<sup>\*\*</sup>Present address: Department of Chemistry, University of California, Riverside, California 9521, U.S.A.



His studies involved a number of properties known to be important factors in biological activity. In this approach, one deduces from a given molecular structure, using some predetermined protocol, a value for the index which is subsequently used in seeking correlations. Several such indices have been suggested in the literature [8–11]. An index based upon discrimination of bond types depending on the number of nearest neighbors, the connectivity index [12], and its modifications [13] has found some application in quantitative structure–activity studies.

In the ongoing research, efforts have been made to generalize the rules governing the construction of topological indices, so that broader classes of compounds can be adequately treated. As part of this work, the simultaneous use of several indices (in the form of ordered sequences of numbers derived from graph invariants) has been explored. The graph invariants utilized are graph paths, which are particularly convenient for the quantitative structural comparisons desired. Although the details of this method have been fully described elsewhere [14–16], and the utility of the technique for a variety of diverse applications including studies of molecular thermodynamic [14], spectroscopic [17], and activity properties [15, 16] has been demonstrated, for clarity the method will be briefly described here with a simple example.

#### STRUCTURAL SIMILARITY BY THE GRAPH PATH METHOD

A convenient, yet simple example of the method can be provided by consideration of the molecular graphs of the hexane isomers (Fig. 1). Each carbon atom is represented by a node (point of intersection of two lines). Hydrogen atoms are not considered at all, in the simplest form of the encoding procedure. Paths are therefore simply sequences of adjacent bonds (lines) connecting carbon atoms (nodes) with no atom appearing more than once in a sequence. One way to represent a molecule is to enumerate the paths of each length emanating from each carbon atom and sum them in a path length-ordered sequence. The sum of paths of length zero corresponds to the number of carbon atoms (obviously the same for all isomers) and the sum of paths of length one corresponds to the number of C–C bonds (again, the same for all isomers). For paths of length one or greater, one must divide the sum by two to obtain molecular path numbers, because each path is

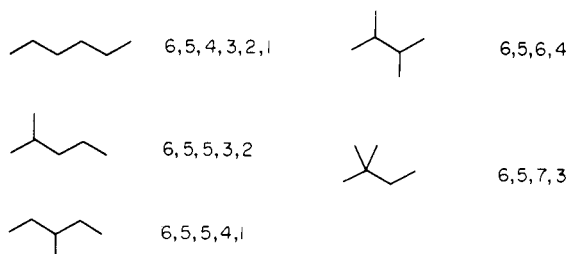


Fig. 1. Molecular graphs of the hexane isomers.

counted twice, once for each end atom. Upon application of the procedure to the hexane isomers, the results of Table 1 are obtained. These five sequences can be used to represent the structures. Furthermore, the differences between the codes can be related to structural differences. This becomes even more apparent when larger structures are considered. Figure 2 illustrates the concept with somewhat more complicated examples. Here, molecular graphs for isomeric dimethylcycloheptane and trimethylcyclohexanes are shown, together with their molecular codes [18]. It can be seen that the molecular codes do reflect the structural similarities (and dissimilarities) of the molecules. The codes for the cycloheptyl systems show the frequent occurrence of the numbers 11 and 12, while the cyclohexyl systems have codes with numbers of 13 or higher appearing frequently. Because these are isomers, all codes begin identically with the numbers 9,9. However, such qualitative observations must be reduced to a systematic and quantitative algorithm if they are to be the basis for quantitative structural comparisons.

As described earlier [18], a suitable quantitative measure of similarity is derived by considering the path numbers as coordinates in an  $n$ -dimensional Euclidean space. Thus, each molecule occupies a point in that space and similarity can be calculated by calculating the distance between points (molecules). Very similar compounds should have graphs which are close together in space and dissimilar compounds should be far apart. Obviously, this basic approach could be modified as appropriate by various normalizations, truncation of the codes, etc. In any event, a quantitative similarity index is easily defined in terms of the Euclidean distance [18]. Equation 1 summarizes these relationships for two hypothetical sequences  $a$  and  $b$ :

$$S^{-1} = D_{ab} = \left[ \sum_i (a_i - b_i)^2 \right]^{1/2} \quad (1)$$

Singularity (i.e.,  $D = 0$  or  $S = \infty$ ) would signify identical sequences.

Further extension of the approach could include the introduction of weighting factors (to make possible incorporation of heteroatom and direct bond-type information). In fact, an algorithm has been published for the latter procedure [19], together with a BASIC program for implementing it. An algorithm and program for the simpler approach described above has also

TABLE 1

Path enumerations for hexane isomers

	Path length					
	0	1	2	3	4	5
n-Hexane	6	5	4	3	2	1
2-Methylhexane	6	5	5	3	2	
3-Methylhexane	6	5	5	4	1	
2,3-Dimethylhexane	6	5	6	4		
2,2-Dimethylhexane	6	5	7	3		

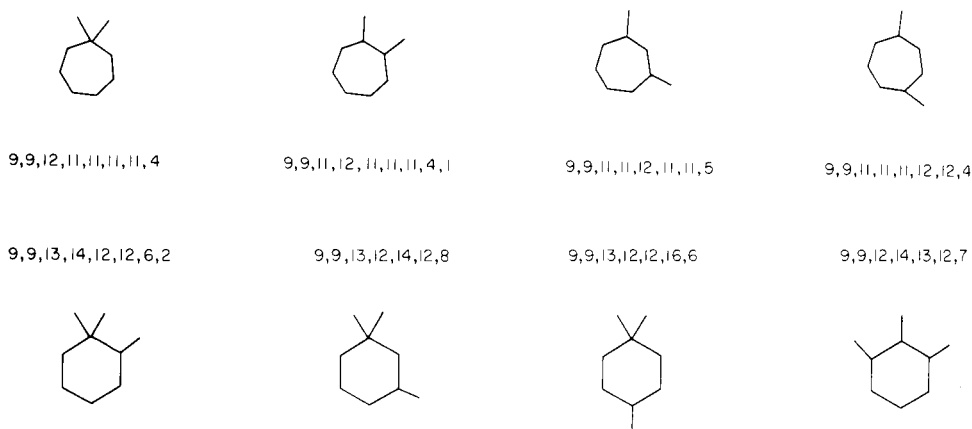


Fig. 2. Molecular graphs for isomeric dimethylcycloheptanes and trimethylcyclohexanes, with molecular codes.

been published [20]. However, in the work described here, the shortcomings of not including these additional kinds of information have been avoided by judicious choice of compounds for study, where heteroatom and bond-type differences are minimized by holding those structural factors approximately constant.

#### STRUCTURAL SIMILARITY AND BIOLOGICAL PROPERTIES

It has for some time been generally accepted that structurally similar compounds should show similarity in their biological activity. One of the aims of the present work is to put this intuitively reasonable concept on a more quantitative basis. As described above, the approach is to assess quantitative similarity using the graph-path method and then to examine relationships of various kinds of biological behavior to the derived similarity indices. An initial attempt in this direction [18] was an examination of a set of computer-generated hypothetical monoterpenes [21] in order to predict which might be the best candidates to be sought in nature, based upon their similarity to known naturally occurring substances. In that study, it was found that the naturally occurring terpenes showed great similarity amongst themselves, a circumstance to be expected on chemical taxonomic grounds. This result is partial verification of the plausibility of the method, although it does not clearly prove its worth.

Accordingly, the general methodology has been further illustrated in two separate additional studies. In the first [15], a series of benzomorphan narcotic analgesics was examined with respect to their biological test data consisting of  $ED_{50}$  values for hot-plate tests of mice [22], compiled by Kaufman et al. [23]; it was found to be possible to select from among many potentially active and closely similar structures those which were indeed the

most active. A second study [16] was devoted to examination of the cerebral dopamine agonist properties of eighteen *N*-substituted 2-aminotetralins and prediction of expected activity for an additional ten hypothetical 2-aminotetralin derivatives. Activity data utilized for that work was based on behavioral effects in mice (sniffing, gnawing, and hyperactivity) produced when the drugs were administered [24–26]. Once more, the similarity indexes were closer together among the most active compounds whereas the inactive compounds were found, in general, to be substantially dissimilar from the active ones. It was concluded [16] that the approach may have application whenever one has several standards with which other substances can be compared for structural similarity. It was also apparent, as in the previous study of benzomorphanes, that this is an economical technique for ranking candidate compounds prior to screening for a desired activity or property.

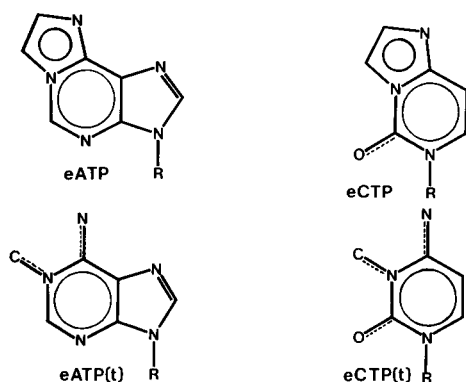
Here, additional results of application of the graph-theoretic method to two structure–activity/reactivity studies of current interest are described. The two applications are quite different, but further illustrate the promise of the technique. The first of these is a study of hexokinase reaction rates in the presence of various nucleotides. The second is an examination of the activity of enkephalin analogs.

#### *Hexokinase reaction rates*

An important enzyme in the metabolism and ultimate conversion of sugars to energy is hexokinase. Although for some time there was disagreement concerning the details of its reaction mechanism, it is now generally agreed that it involves a random addition of substrates, a nucleoside triphosphate, and a hexose [27–30]. Although adenosine triphosphate (ATP) is the most commonly used nucleotide, the enzyme is capable of utilizing a variety of other nucleoside triphosphates [30]. Studies on the effects on  $K_m$  and  $V_{max}$  for these reactions caused by changes in the structure of the nitrogen base have been reported [30, 31]. Thus, the application of the graph-theoretic approach described above seemed attractive.

It is known [30] that the size of the parent base is important to the ability of the nucleotide to function as a phosphate donor in the hexokinase reaction. Furthermore, the 100-fold range of both  $K_m$  and  $V_{max}$  for purine substrates implies selectivities of the enzyme based on the functional groups of the rings forming the nitrogen bases, because ribose triphosphate is common to all these substrates.

For graph-theoretical treatment, a number of simplifications were made. Structures were first simplified by replacing the ribose-5'-triphosphate (common to all the nucleotides) with the pseudatom R. For similarity comparisons, only atoms directly attached to the parent ring were considered. Thus, for ethenoadenosine (eATP) and ethenocytidine (eCTP), the structures were truncated as shown below (eATP(t) and eCTP(t), respectively).



As in the benzomorphan and dopamine agonist studies, atomic identity was not explicitly used. Molecular graph paths were calculated as described earlier and a similarity matrix was constructed. For the purines, ATP, which has the highest  $V_{\max}$ , was chosen as the reference compound, and for the pyrimidines, 5-bromouridine (formally identical to thymidine, since atom identities are ignored) served as the reference. Kinetic parameters,  $V_{\max}$  and  $K_m$ , were obtained from standard biochemical assays of hexokinase activity. Those data are summarized in Table 2.

Examination of these results shows that, with one exception (eATP, actually a tricyclic base) all of the purine triphosphates have an effect on  $V_{\max}$  directly proportional to their graph path similarity to ATP. For the pyrimidines an analogous result is found. Again there is a single exception which probably results from the formal identity in the present treatment of different atoms. A noteworthy outcome is the absence of any such parallelism for the  $K_m$  values in either series. One possible interpretation of these facts

TABLE 2

Kinetic parameters for nucleoside triphosphate interaction with yeast hexokinase

Substrate	$V_{\max}$ ( $\mu\text{mol min}^{-1} \text{mg}^{-1}$ )	$K_m$ (mM)	$S \times 10^3$
<i>Purines</i>			
Adenosine triphosphate (ATP)	18.60	0.35	—
Ethenoadenosine triphosphate	13.80	0.23	6
Inosine triphosphate	9.95	9.13	28
8-Bromoadenosine triphosphate	5.95	0.11	11
Guanosine triphosphate	3.52	4.35	9
Xanthosine triphosphate	0.36	1.00	6
<i>Pyrimidines</i>			
Ethenocytidine triphosphate	9.85	0.52	—
5-Bromouridine triphosphate	6.70	3.03	37
Uridine triphosphate	2.17	6.31	30
Cytidine triphosphate	1.44	10.60	21
Thymidine triphosphate	0.20	2.67	37

is that the catalytic constant (reflected in  $V_{\text{max}}$ ) is proportional to the rate of the slow step of the reaction, while  $K_m$  is not. This suggests that  $K_m$  is dominated by a rapid equilibrium between free enzyme and nucleotide. Thus  $K_m = K_{\text{dis}}$  for the enzyme complex and would be controlled by thermodynamic rather than kinetic factors. Other studies in this laboratory [30] have validated the identity of  $K_m$  and  $K_{\text{dis}}$ . Thus, the graph path approach in the present study focused attention on one of the key details of the reaction mechanism.

### *Opiate-like activity of enkephalins*

Enkephalins are naturally-occurring molecules which have opiate-like analgesic effects. They have been shown to react with morphine-type receptors in the brain, intestine, and several other organs. Since their discovery, a large number of these hormones and analogs have been synthesized and tested. The choice of compounds synthesized has been based on classical views of structure—activity models, generally involving minor modifications of known active substances. In the present examination of enkephalins, graph path methods were used to predict the activity of proposed enkephalins by comparison with known materials possessing opiate activity. As in the other studies reported above, the molecular graph path sequences were computed, followed by the similarity indices (relative to the selected standards). Table 3 lists the compounds compared and their relative activities.

TABLE 3

The enkephalins (1 and 2) and other structures tested for opiate activity

	Compound	Relative activity <sup>a, b</sup>	Ref.
1	Tyr-Gly-Gly-Phe-Met	1	32
2	Tyr-Gly-Gly-Phe-Leu	0.5	32
3	Tyr-DAla-Gly-Phe-Met	100	32
4	Tyr-Met-Gly-Phe-Pro-amide	1,500	32
5	Tyr-DAla-Gly-Phe-Met-ol	1,600	32
6	Tyr-DAla-Gly-Phe-Met(s)-ol	9,600	32
7	Tyr-DAla-Gly-Phe*-Met(s)-ol	28,800	32
8	Tyr-Gly	Inactive	33
9	Tyr-Gly-Gly	Inactive	33
10	N-Acetyl Tyr-Gly-Gly-Phe-Met	Inactive	33
11	Phe-Gly-Gly-Phe-Met	0.002	34
12	Tyr-Gly-Gly-Tyr-Met	0.003	34
13	Tyr-Gly-Ala-Phe-Met	0.04	33
14	(OMe) Tyr-Gly-Gly-Phe-Met	0.17	33
15	Arg-Tyr-Gly-Gly-Phe-Met	0.25	35
16	Tyr-Gly-Gly-Phe-NorLeu	0.45	35
17	Tyr-Gly-Gly-Phe-Met-Nethyl	2.46	33
18	Tyr-Gly-Gly-Phe-Met-NH <sub>2</sub>	3.02	33

<sup>a</sup> Abbreviations: Met(s) = methionine sulfoxide; Phe\* = *N*-methylphenylalanine; ol = alcohol substitution on carboxyl terminus. <sup>b</sup> All values normalized to methionine enkephalin.

The seven most commonly studied enkephalins and analogs were arbitrarily chosen as standards. The similarities of each of the remaining eleven compounds with respect to these seven are compiled in Table 4. These values were examined to test the hypothesis that compounds of comparable activity would have larger values of  $S$  with respect to the standards. The results lend some credence to this idea. Specifically, it is obvious that the two inactive compounds (8 and 9) are quite dissimilar from all of the standards. This is as expected. However, the third inactive compound (10) is not discernibly different from the standards, except for the most active standard. Yet, compound 15, which has 0.25 times the relative activity of Met-enkephalin (compound 1) appears almost as dissimilar as the two inactive compounds. The material of highest activity amongst the eleven examined is clearly far more similar with the standards (with the exception of compounds 4 and 7, which yield low values of  $S$  for all compounds). These data do suggest that 7, by far the most active substance considered, may well owe its activity to different structural features than those responsible for the activity of the others. Thus, although the graph path similarity measure is not definitive (for the reasons discussed earlier in the paper) it does appear to offer some useful insights.

One of the potential values of this method is the possibility of using the technique to screen large numbers of computer-generated structures in a search for those sufficiently promising for synthesis. In the present research, the eight standards were used as illustrated above for some typical structures to screen over 3,000 pentapeptides incorporating an *N*-terminal tyrosine and a phenylalanine in the fourth position. From among those screened, 60 compounds were predicted to have activity in the range exhibited by the standards.

TABLE 4

Similarity matrix for standard vs. tested enkephalin analogs<sup>a, b</sup>

Analog	Standards						
	1	2	3	4	5	6	7
8	4	4	4	3	4	4	2
9	4	4	4	3	4	4	2
10	41	40	43	16	42	36	4
11	27	26	30	12	33	32	4
12	27	28	24	18	21	20	4
13	54	49	68	14	79	69	4
14	33	31	28	17	26	26	4
15	6	6	7	5	7	7	3
16	27	26	29	17	28	27	4
17	31	30	27	20	24	24	4
18	132	96	101	15	88	67	4

<sup>a</sup>Compounds are those listed in Table 3. <sup>b</sup>Similarity values listed were computed using eqn. (1) and have been multiplied by 100.

Another interesting fact emerged from examination of the screening results. None of the predicted active materials showed marked similarity with compound 7, the most active among those studied. However, if only similarity with 7 is considered, the six compounds significantly closer to 7 possess the common structural feature of a C-terminal proline. This fact, together with the interesting observation that these same six are also the most similar to compound 4 (a proline-containing standard), suggests that 4 and 7 may well owe their activity to different factors than those responsible for activity in the remaining standards. A possible explanation for changed behavior may be the restriction of rotation about the  $\alpha$ -C—N bond imposed by the proline moiety. The synthesis of these six compounds is being studied.

The authors gratefully acknowledge support from the National Science Foundation (Grant CHE-79-10263 and Grant PCM-78-12134). S.M.S. was supported by a National Cancer Institute Research Career Development Award, CA-00628-01.

#### REFERENCES

- 1 C. Hansch, *Acc. Chem. Res.*, 2 (1969) 232.
- 2 S. M. Free and J. W. Wilson, *J. Med. Chem.*, 7 (1964) 395.
- 3 L. J. Soltzberg and C. L. Wilkins, *J. Am. Chem. Soc.*, 99 (1977) 439.
- 4 F. Peradejordi, A. N. Martin, and A. Cammarata, *J. Pharm. Chem.*, (1971) 576.
- 5 A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, 97 (1975) 182.
- 6 B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, 96 (1974) 916.
- 7 H. Wiener, *J. Am. Chem. Soc.*, 69 (1947) 2636.
- 8 H. Hosoya, *Bull. Chem. Soc. Jpn.*, 44 (1971) 2332.
- 9 L. Lovasz and J. Pelikan, *Period. Math. Hung.*, 3 (1973) 175.
- 10 M. Gordon and G. R. Scantlebury, *Trans. Faraday Soc.*, 60 (1964) 605.
- 11 D. Bonchev and N. Trinajstić, *J. Chem. Phys.*, 67 (1977) 4517.
- 12 M. Randić, *J. Am. Chem. Soc.*, 97 (1975) 6609.
- 13 L. B. Kier and L. L. Hall, *Molecular Connectivity in Chemistry and Drug Research, Medicinal Chemistry*, Vol. 14, Academic Press, New York, 1976.
- 14 M. Randić and C. L. Wilkins, *J. Phys. Chem.*, 83 (1979) 1525.
- 15 M. Randić and C. L. Wilkins, *Int. J. Quantum. Chem., Quantum Biol. Symp.*, 6 (1979) 55.
- 16 C. L. Wilkins and M. Randić, *Theor. Chim. Acta*, 58 (1980) 45.
- 17 M. Randić, *J. Magn. Reson.*, 39 (1980) 431.
- 18 M. Randić and C. L. Wilkins, *J. Chem. Inf. Comp. Sci.*, 19 (1979) 31.
- 19 M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. Chem.*, 4 (1979) 27.
- 20 M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. Chem.*, 3 (1979) 5.
- 21 D. H. Smith and R. E. Carhart, *Tetrahedron*, 32 (1976) 2513.
- 22 N. B. Eddy and D. Leimbach, *J. Pharmacol. Exp. Ther.*, 107 (1953) 385.
- 23 S. J. Kaufman, A. E. Jacobson, and W. F. Raub, *J. Chem. Doc.*, 10 (1970) 248.
- 24 J. D. McDermed, G. M. McKenzie, and A. P. Phillips, *J. Med. Chem.*, 18 (1975) 362.
- 25 J. G. Cannon, J. C. Kim, M. A. Aleem, and J. P. Long, *J. Med. Chem.*, 15 (1972) 348.
- 26 J. G. Cannon, T. Lee, and H. D. Goldman, *J. Med. Chem.*, 20 (1977) 1111.
- 27 J. D. Lueck and H. J. Fromm, *J. Biol. Chem.*, 249 (1974) 1341.
- 28 K. D. Danenberg, and W. W. Cleland, *Biochemistry*, 14 (1975) 28.
- 29 T. J. C. Higgins and J. S. Easterby, *Eur. J. Biochem.*, 65 (1976) 513.
- 30 L. J. Dorgan and S. M. Schuster, *Arch. Biochem. Biophys.*, 207 (1981) 165.
- 31 D. C. Hohnadel and C. Cooper, *Eur. J. Biochem.*, 31 (1972) 180.
- 32 S. H. Synder, *Chem. Eng. News*, 55 (1977) 26.
- 33 N. Ling and R. Guillemin, *Proc. Natl. Acad. Sci. U.S.A.*, 73 (1976) 3308.
- 34 B. A. Morgan, C. F. C. Smith, A. A. Waterfield, Hughes, Jr. and H. W. Kosterlitz, *J. Pharm. Pharmacol.*, 28 (1976) 660.
- 35 T. T. Pham Chau, M. Lujan, W. L. Dewey, R. J. Freer, A. R. Day and L. S. Harris, *Pharmacologist*, 18 (1976) 120.



## A GRAPH THEORY DATA BASE FOR STORAGE OF CHEMICAL STRUCTURES ORGANIZED BY THE BLOCK-CUTPOINT TREE TECHNIQUE

YUZURU FUJIWARA\* and TAKASHI NAKAYAMA

*Institute of Information Sciences and Electronics, University of Tsukuba, Sakura-mura, Niihari-gun, Ibaraki 305 (Japan)*

(Received 23rd January 1981)

### SUMMARY

The method presented for organizing a data base according to graph theory, is based on representation of chemical structures in terms of BCT (block-cutpoint tree). It is useful for quick substructure searches and is convenient for structure generation. The data base consists of four files: a master file, a bit sequence file of fixed length records which gives block components of compounds, a BCT file which gives the BCT structures of compounds, and a block file which specifies the blocks. These files are organized recursively and hierarchally, which simplifies the processing of structural information on compounds.

Information about chemical structures is essential in many applications, such as substructure searches, structure elucidation from spectral data, drug design, and planning of syntheses. A typical data base of chemical structures has the data in the form of connection tables or adjacency matrices. Data bases of this type are independent of applications, and therefore universal, but are not necessarily efficient for structural data processing such as access to data through a substructure as a key or for searching a series of compounds with similar structures.

One of the major problems in computer handling of chemical structures is the explosive increase of time and space complexities caused by increase in the number of atomic combinations. Application systems should be designed to reduce these complexities. The solutions lie in two directions: improvement of processing algorithms and improvements in methods of representing chemical structures. Many efforts have been made in both directions, and a typical method in the second direction is the screen system for substructure search [1-3]. The chemical structure data base used here is organized by the BCT (block-cutpoint tree) system [4] and is another attempt to improve representation of chemical structures. This data base is universal and independent of particular applications. It is constructed so that time and space complexities may be substantially reduced, for effective use of information about chemical structures.

In a chemical structure data base (CSDB), the representation of structures should not use codes that simply identify compounds and require an algorithm to assign unique and unambiguous codes. These codes are of little use from the data-base viewpoint unless they are recursive and suitable for screening. The present data base has taken this into account and does not deal with atoms directly, but uses a block consisting of several atoms as a meaningful operational unit. A block is an intermediate concept between atoms and molecules, and corresponds to a BCT in the present representation. A superblock composed of simple blocks is introduced as a high-order operational unit which is easily defined and modified for flexible and effective processing of structures.

## REPRESENTATION OF CHEMICAL STRUCTURES

In recent work, a hierarchic representation of chemical structures by means of the BCT system was presented [5]. In this representation, an intermediate representation is generated in the form of a tree whose nodes consist of blocks and cutpoints, where blocks (doubly-connected components of a graph) correspond to ring systems or linear systems of two vertices. The strict representation is described by connectivity information between blocks and block structures in the block dictionary. A brief description of BCT representation of chemical structures is given below:

### *Block-cutpoint tree*

Chemical structures are regarded as graphs in this method, and atoms and bonds correspond to vertices and edges, respectively. A vertex of a connected graph  $G$  is called a cutpoint if its removal causes disconnection of  $G$ , and a block is a maximal subgraph of  $G$  which contains no cutpoints. Let  $\{B_i\}$  be the set of blocks of  $G$ , and  $\{c_j\}$  be the set of cutpoints of  $G$ ; then a block-cutpoint graph of  $G$  is defined as a graph whose vertex set is  $\{B_i\} \cup \{c_j\}$  and whose connectivity is defined between  $\{B_i\}$  and  $\{c_j\}$ , where  $c_j$  is adjacent to  $B_i$  only if  $c_j$  is one of the vertices of block  $B_i$ . The block-cutpoint graph is always a tree, so it is called a block-cutpoint tree (hereafter a BCT). Cutpoints and blocks of a graph and its BCT are shown in Fig. 1. Black nodes in Fig. 1 are cutpoints.

### *BCT representation of chemical structures*

The algorithm which finds the blocks and cutpoints of a graph has been described [5]; this was based on two other algorithms [6, 7]. A BCT is constructed easily from those blocks and cutpoints by a program called BCTGEN which also finds the blocks and cutpoints. Input data for BCTGEN are adjacency matrices or connection tables of graphs. Given the data for a chemical structure, BCTGEN generates a BCT as its intermediate structure, and describes it in canonical form where the center of the structure is the root of the tree. This was implemented by Edmond's algorithm [8]. Block

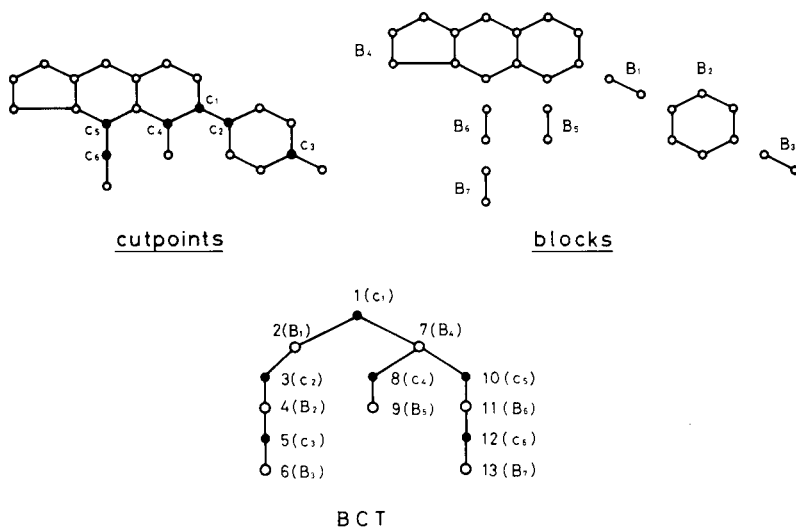


Fig. 1. Cutpoints, blocks and BCT of a graph.

nodes (a member of the vertex set of a BCT can be called a node, so that there are block nodes and cutpoint nodes in a BCT) are given identification numbers by BCTGEN, and their internal structures are available in the block dictionary (see below).

Further information about the connectivity of the constituent blocks is needed in order to describe a chemical structure completely. This is expressed by a connectivity matrix which represents the connectivity among blocks: a row and a column of connectivity matrix represent the constituent blocks  $B_1, \dots, B_m$ . The element of the connectivity matrix  $c_{ji}$  represents the identification number of the cutpoint in  $B_i$  which connects  $B_i$  with  $B_j$  if  $B_i$  is connected with  $B_j$ , otherwise  $c_{ij}$  is zero. The element  $c_{ij}$  also represents the connectivity of block  $B_j$  in relation to block  $B_i$ . Thus the connectivity matrix is not symmetrical (usually  $c_{ij} \neq c_{ji}$ ). Morgan's algorithm [9] is used as the numbering scheme for the vertices of a block. The way of selecting cutpoints in a block is not always unique, and so the following method for specifying cutpoints is used. Suppose that  $\{p_1, \dots, p_k\}$  is the set of cutpoints in the block, where  $p_i$  is the identification number. The set  $\{p_1, \dots, p_k\}$  is selected so that it gives the minimal value when the sequence of any permutation of  $p_1, \dots, p_k$  is evaluated from the block dictionary. Further, the following rule is applied because equivalent vertices (vertices belonging to the same orbit) can be selected as cutpoints at the same time. The assignment order of cutpoints in a given block is determined in relation to the blocks which connect with the given block. This means that cutpoints are assigned in depth-first order to each connecting vertex of the adjacent block node in the BCT canonical form, and a smaller number is assigned to a prior (i.e., deeper) vertex in an orbit. Figure 2 shows an example of cutpoint assignment for vertices of blocks.

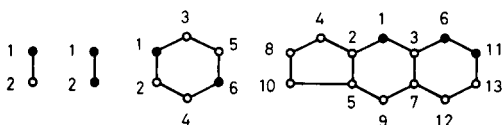


Fig. 2. Example of cutpoint assignment for vertices of blocks.

The BCT representation of the structure in Fig. 1 is shown in Table 1. The tree code consists of four lines. The first line indicates the node number assigned in depth-first order in the BCT canonical form. The second line is the tree code computed by Edmonds' algorithm, and the number in each column represents the number of descendants (including itself) of the node which is specified by the number above it in the first line. The third line shows the identifiers of the block nodes specified by the first line. These are key values of the block dictionary, but in this example alphabetic symbols have been used; in practice they are registry numbers. The connectivity matrix is shown in the lower part of Table 1. Row/column numbers correspond to the numbers of blocks indicated in the first line of the tree code. The numbering of vertices in a block is shown in Fig. 2. Therefore the first row of the connectivity matrix indicates that block 2 is connected with block 4 at vertex 1, and with block 7 at vertex 2. The second row shows that block 4 is connected with block 2 at vertex 1, and with block 6 at vertex 4. Rows 3–7 are interpreted similarly. Actually the connectivity

TABLE 1

BCT representation of a chemical structure

*BCT code*

1	2	3	4	5	6	7	8	9	10	11	12	13														
13	5	4	3	2	1	7	2	1	4	3	2	1														
	<i>a</i>		<i>b</i>		<i>c</i>	<i>d</i>		<i>c</i>		<i>a</i>		<i>c</i>														
4	<u>4</u>	<u>1</u>	<u>7</u>	<u>2</u>	<u>4</u>	<u>2</u>	<u>1</u>	<u>6</u>	<u>4</u>	<u>2</u>	<u>11</u>	<u>9</u>	<u>6</u>	<u>11</u>	<u>1</u>	<u>2</u>	<u>7</u>	<u>1</u>	<u>4</u>	<u>7</u>	<u>1</u>	<u>13</u>	<u>2</u>	<u>2</u>	<u>11</u>	<u>1</u>

*Connectivity matrix*

	2	4	6	7	9	11	13
2	0	1	0	2	0	0	0
4	1	0	4	0	0	0	0
6	0	1	0	0	0	0	0
7	11	0	0	0	6	1	0
9	0	0	0	1	0	0	0
11	0	0	0	1	0	0	2
13	0	0	0	0	0	1	0

matrix is linked to the tree code in a list form, as shown in the fourth line of the BCT code in Table 1.

The hierarchal representation of chemical structures is achieved through the use of BCT. This representation of chemical structures uses records of the BCT type, which are intermediate structure representations, and of the BCF type, which represent the block components of a chemical structure and neglect connectivity among blocks. These two types of record make it possible to access the data base flexibly and rapidly.

#### CONSTRUCTION OF THE CHEMICAL STRUCTURE DATA BASE

The layout of a chemical structure data base (CSDB) constructed on the basis of the BCT representation of chemical structures is shown in Fig. 3. The network configuration allows a loop (recursive reference) but the access path from a compound through an intermediate representation in block terms to internal structures of blocks reflects the hierarchy of the BCT representation. The characteristic feature of CSDB is the network structure incorporating both hierarchy and recursiveness. Types of record in the CSDB are explained below.

##### *Types of record*

*Master file.* The master file has all the attributes of the compounds except structural data: CAS registry number, compound name, molecular formula, molecular weight and identification number are included as key items. New items can be added in order to record physical properties of various kinds as attributes of compounds.

*BCT file.* The BCT file is one of two files containing structural data for compounds. This file has the records of the BCT representation of chemical structures, where the internal structures of constituent blocks are available in the

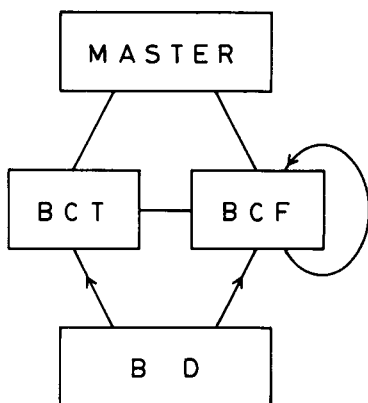


Fig. 3. Block diagram of the chemical structure data base.

block dictionary (BD). The key item of the BCT record is common to the master and BCF files. Cross-reference between these files is accomplished by that key item, and the lines between these three types of record in Fig. 3 show it.

The upper part of Table 1 shows the BCT representation of the structure in Fig. 1. The fourth line represents the connectivity matrix; a number without underlining indicates the length of the following underlined sub-sequence. Each underlined sub-sequence corresponds to a row of the connectivity matrix shown in the lower part of Table 1 in row number order. For example, sub-sequence 4172 means that block 2 is connected with block 4 at vertex 1, and with block 7 at vertex 2. The other lines were explained in the previous section. The correspondence of BCT to BD is  $n:1$ .

*Block component file.* The block component file (BCF) contains the block components of chemical structures; it indicates which blocks are contained in a compound, and so can be regarded as a kind of fragment file. The BCF format is shown in Fig. 4. The record length is fixed at 1024 bits excluding the identifier area. A bit position or a group of bit positions corresponds to identifiers of constituent blocks of a compound.

The record is divided into two areas: a simple block area and a superblock area. A simple block is a block defined graph-theoretically as described before; the term "block" without the qualifier "simple" is used in the following discussion to mean a simple block. A superblock is a subgraph which consists of connected blocks, and is used mainly for representing substructures which contain acyclic parts (the BCT representation is not always advantageous for representing acyclic substructures). Each area is further subdivided into two parts: core and extension. The core contains blocks/superblocks that occur with high frequency in compounds, and each bit position in a core corresponds to an identifier of a block/superblock; 416 blocks and 160 superblocks are assigned to cores of blocks and superblocks, respectively.

The extension contains blocks/superblocks which occur less frequently in compounds. It consists of a modifier part of 64 bits and a bit position part of 160 bits. A block/superblock in an extension is expressed by a group of

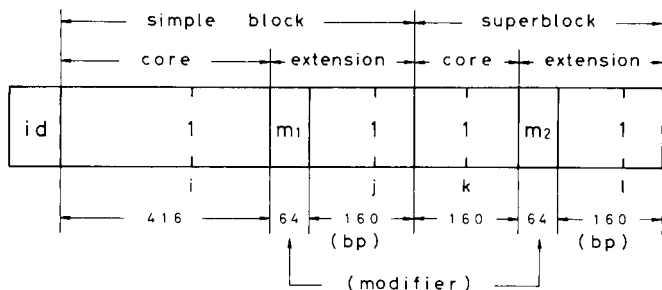


Fig. 4. Format of the block component file.

bits  $(m,p)$ , where  $m$  is the value of a modifier part and  $p$  is an on-bit position of a bit position. The term  $(m,p)$  means that the record is stored logically as the  $p$ th record in the  $m$ th page of a BD/SBD file. In Fig. 4, the record indicates that the two blocks  $i$  and  $(m_1,j)$  and the two superblocks  $k$  and  $(m_2,l)$  are the components of the structure.

A modifier is defined so that it reflects the chemical and topological features of a chemical structure. The contents of the modifiers are shown in Table 2. If more than one bit is set in the bit position part, the logical sum of the respective modifiers is stored in the modifier part:  $m = m_1 \vee m_2 \vee \dots \vee m_k$  where  $m_i$  is a modifier corresponding to a single on-bit ( $i = 1, \dots, k$ ). In this case, the extension part becomes  $(m, p_1, \dots, p_k)$ . This would seem to indicate that blocks/superblocks  $p_1, \dots, p_k$  are stored in the same page, but this conflict is resolved by referring to a thesaurus of modifiers. If the number of blocks/superblocks of the same modifier exceeds 160, the 7th and 8th bytes of the modifier are used for extending page numbers. Their value (modulus 160) is taken as the least significant part of the page number. Therefore,  $2^{16} \times 160$  blocks/superblocks can be expressed in a modifier. An example of a BCF record is shown in Fig. 5. Blocks  $B_1, B_2, B_3, B_5$  and  $B_6$  are members of a core part; block  $B_4$  is a member of an extension part and is also superblock  $SB_1$ . Modifiers  $m_1$  and  $m_2$  are as follows (bytes 1–6):

```

m1 = 00100000 00010000 10000000
      00000001 10000000 00001000
m2 = 01001000 01001000 10000000
      00001001 10000000 00001000

```

BCT and BCF are indexed sequential files. A modifier of the same format as that of superblocks is computed for each compound, and is used for indexing the storage page for compounds. The storage page consists of 160 records corresponding to the BCF format. A page suffix is inserted when the number of compounds of the same modifier exceeds 160. The suffix is also put in the modifier itself (the value in the 7th and 8th byte of the modifier, as noted earlier).

*Block dictionary.* The block dictionary (BD) is a file of the contents of blocks (structure code) which provides the elements for describing BCT/BCF records. The block code is in the form of a connection table. The numbering of the atoms is based on Morgan's algorithm. The constituent unit of the BD file is a page of 160 records corresponding to the BCF record format. The procedure for registering blocks is as follows: the modifier is computed for a given block, then the logical page for the modifier is consulted; the block is registered if it is not found. If the page itself has not yet been set up, the block is registered after it is set up. If the number of registered blocks of the page exceeds 160, a new page is added with a suffix which is put in the 7th and 8th byte of the modifier. Access to the file is managed by the modifier thesaurus which is also a directory to the BD file.

*Superblock dictionary.* The superblock dictionary (SBD) gives the contents (structures) of the superblocks that appear in BCF records. In this sense,

TABLE 2

Contents of the simple block and superblock modifiers

SIMPLE BLOCK MODIFIER				SUPERBLOCK MODIFIER						
Byte	Bit			Byte	Bit					
1	1	No. of rings	= 1	1	1	No. of acyclic block types	= 1			
	2		= 2		2		= 2			
	3		= 3		3		= 3			
	4		= 4		4		> 4			
	5		= 5		5		No. of cyclic block types	= 1		
	6		= 6		6			= 2		
	7		= 7		7			= 3		
	8		≥ 8		8			> 4		
2	1	Ring type	3-ring	2	1	No. of acyclic blocks	= 1			
	2		4-ring		2		= 2			
	3		5-ring		3		= 3			
	4		6-ring		4		> 4			
	5		7-ring		5		No. of cyclic blocks	= 1		
	6		8-ring		6			= 2		
	7		9-ring		7			= 3		
	8		≥ 10-ring		8			> 4		
3	1	Core/extension	= 0/1	3	1	Core/skeleton	= 0/			
	2 <sup>a</sup>				2-8 BCT skeleton id. no.					
4	1	No. of double bonds	= 1	4	1-8	BCT skeleton id. no.				
	2		= 2							
	3		= 3				5	1	No. of N atoms	= 1
	4		≥ 4							2
	5	No. of triple bonds	= 1	3	= 3					
	6		≥ 2	4	= 4					
	7	No. of aromatic bonds	= 0	5	= 5					
	8		≥ 1	6	> 6					
5	1	No. of N atoms	= 1	6	1					
	2		= 2					2	= 3	
	3		= 3					3	= 4	
	4		= 4					4	= 5	
	5	= 5	5	> 6						
	6	≥ 6	6	No. of O atoms	= 1					
	7	No. of O atoms	= 1		7	= 2				
	8		= 2		8	= 3				
						= 4				
					= 5					
6	1	No. of S atoms	= 1	6	1					
	2		= 2					2	= 3	
	3		= 3					3	= 4	
	4		≥ 4					4	= 5	
	5	= 5	5	> 6						
	6	No. of S atoms	= 1	6	= 1					
	7		= 2	7	= 2					
	8		= 3	8	= 3					
	≥ 4			> 4						

<sup>a</sup>Bits 2-8 here are not used.



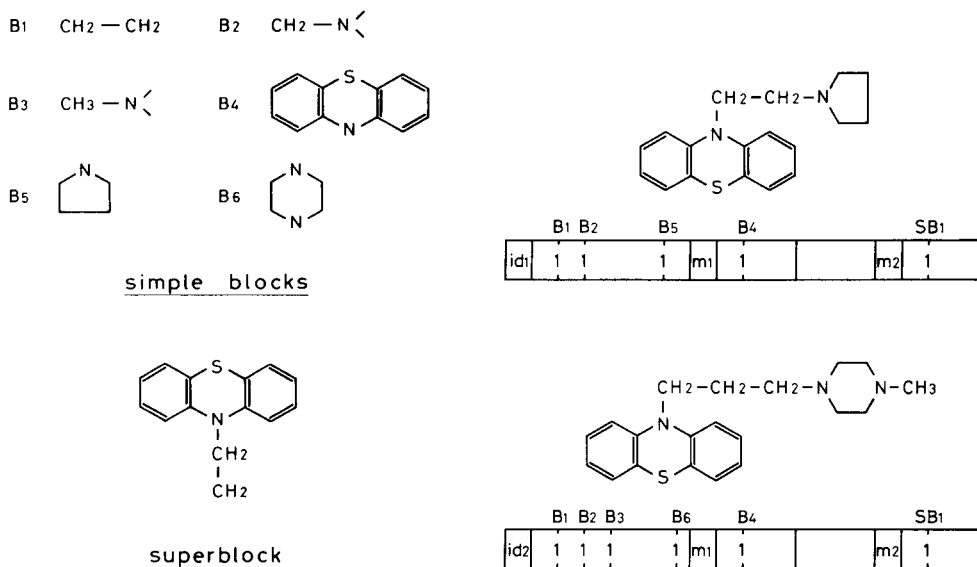


Fig. 5. Example of records in the block component file.

the SBD plays the same role as the BD in the CSDB, but it is not strictly an independent file. Rather, it is defined as another cycle of the BCT/BCF file, because the structures of superblocks are equivalent to ordinary compounds. The SBD and BCT/BCF hold a record in common, and the key item (the identifier area in Fig. 3) of the record contains flag bits for identifying the membership of the record: bit 1, set if the record is a member of BCT/BCF; bit 2, set if the record is a member of SBD; bits 3–32, the registry number of the record.

The number of entries in the SBD is kept at the same order of magnitude as the number of BD entries. As the main reason for using superblocks is to simplify the description of search profiles, the selection of SBD entries is based on a statistical evaluation of search profiles. Access to the SBD is managed by the modifier thesaurus which is also a directory to the SBD file.

### Conclusion

A chemical structure data base which makes it easy to access substructures in various aspects can be formed by using blocks and superblocks as intermediates describing units for representing chemical structures. In particular, the representation by bit sequences makes it easy to abstract specified substructures, which is an operation corresponding to imaging for attributes of relational data-base models. Any combination of Boolean logic operations, including negation, is easily executed. The Markush formulae, which are important with regard to patent information, can also be handled by BCF.

The present method for organizing a CSDB is useful for quick substructure search. When the query structure is represented in BCT form, the character-

istic of the CSDB is fully displayed for substructure search as described below; the query structure is input in the form of a connection table.

(1) *Search pattern making.* The BCT representation of the query structure and a structural description with the same format as the BCF record are made from the input connection table.

(2) *BCF search.* If the query structure is registered in SBD as a superblock, the bit pattern representing it is searched in BCF. If the query structure is not a member of SBD, the bit pattern representing its block components is sought as a screening step, and then the connectivity among blocks is checked by consulting the BCT file.

This CSDB is suitable for structure inference from spectral data, molecular design, and other application systems.

## REFERENCES

- 1 M. F. Lynch, in J. E. Ash and E. Hyde, *Chemical Information Systems*, Ellis Horwood, 1975, p. 177.
- 2 M. Milne, D. Lefkowitz, H. Hill and R. Powers, *J. Chem. Doc.*, 12 (1972) 183.
- 3 W. Graf, H. K. Kaindl, H. Kniss, B. Schmidt and R. Warszawski, *J. Chem. Inf. Comput. Sci.*, 19 (1979) 51.
- 4 F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.
- 5 T. Nakayama and Y. Fujiwara, *J. Chem. Inf. Comput. Sci.*, 20 (1980) 23.
- 6 I. C. Ross and F. Harary, *Manag. Sci.*, 1 (1955) 251.
- 7 A. V. Aho, J. E. Hopcroft and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- 8 R. G. Busacker and T. L. Saaty, *Finite Graphs and Networks: An Introduction with Applications*, McGraw-Hill, New York, 1965.
- 9 H. L. Morgan, *J. Chem. Doc.*, 5 (1965) 107.

## DEVELOPMENT OF A GRAPHIC PROGRAM FOR QUANTITATIVE DRUG DESIGN

TOSHIYUKI ESAKI

*Department of Applied Chemistry, Faculty of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi 464 (Japan)*

(Received 23rd January 1981)

### SUMMARY

A graphic program for quantitative drug design is described; it is used for the design stage of candidate compounds after a regression equation has been established for estimation of biological potencies in the Hansch-Fujita and Free-Wilson methods. Information flow is controlled by function-key and concomitant alphanumeric-key operations. In the program, substituents are easily introduced into a parent structure, and a perspective molecular model of the resulting derivative is displayed. Conformational changes of a molecule can be followed by rotating atomic groups. The data bank contains 337 possible substituents with 20 physicochemical parameter values and constituent atomic coordinates for each substituent. Examples are shown to illustrate the program.

Studies on drug structure-activity relationships can be made more efficient by computer-assisted techniques [1]. From this standpoint, the process of drug design is generally divided into two steps: (1) generation of new leads; (2) optimization of a previously recognized lead structure by chemical modification. For each step, various techniques have been proposed and applied. The Hansch-Fujita method [2] and the Free-Wilson method [3] have been widely and successfully applied for improvement of efficiency. These methods are distinguished in that the former considers the physicochemical characteristics of substituents (e.g., hydrophobic, steric and electronic) whereas the latter emphasizes the presence of particular substituents. However, both may be reduced mathematically to the multiple regression problem. Thus, in these methods, the minimum number of analogs should be synthesized and biologically tested initially followed by formulation of a regression equation for estimation of biological potencies. Candidate compounds for further synthesis are then designed to promote research efficiency. The procedure can be speeded up if the structures of possible derivatives can be displayed on a cathode-ray screen along with their corresponding biological potency, by making use of conversational and visual display functions of computer graphics. The purpose of this paper is to describe an attempt to develop a graphic program for the stage of candidate design by the Hansch-Fujita and Free-Wilson methods.

## EXPERIMENTAL

The computer program was written in FORTRAN and implemented on the FACOM M-200 computer, the F6233L graphic display unit (refresh type) and the F6202B X-Y plotter at the Nagoya University Computer Center. In the program, a graphic subroutine package (grammar book and manual; 64SP-6010-2; 64SP-6032-1; Fujitsu, Japan) and a subroutine for molecular structure display (NAMOD) [4] are used in drawing figures on the CRT, and for generating perspective diagrams of molecules, respectively.

Cartesian coordinates of constituent atoms of substituents were calculated by means of a modified version of the COORD/1130 program [5].

The QDD program is registered in the program libraries of the Nagoya University Computer Center and the Computation Center of the Institute for Molecular Sciences (Okazaki, Japan). Its listing and manual are available on request.

## RESULTS

The computer graphic program developed (QDD; graphic program for quantitative drug design) improves the step of candidate design after the equations for estimation of biological potencies in the Hansch—Fujita and Free—Wilson methods have been determined.

The program consists of two parts; the statement part with about 1000 lines and the substituent data part with about 2000 lines. The statement consists of a main program and 17 subroutines. The memory size required is 448 kbytes. Users should prepare the coordinates of constituent atoms of the parent molecule, the information on the equation, etc., as input data. The general forms of equations, which users can deal with, are as follows:

$$BA = a(\sum_i X_{Hi})^2 + b(\sum_i X_{Hi}) + c(\sum_i X_{Ei}) + d(\sum_i X_{Si}) + e \quad (1)$$

$$BA = \sum_i a_i X_{Hi} + \sum_i b_i X_{Ei} + \sum_i c_i X_{Si} + \sum_i d_i D_i + e \quad (2)$$

$$BA = \mu + \sum_{i,j} G_{ij} X_{ij} \quad (3)$$

Equations (1) and (2) are for the Hansch—Fujita method, and eqn. (3) for the Free—Wilson method. In these equations,  $BA$  represents the calculated biological potency of the analog in question;  $X_H$ ,  $X_E$  and  $X_S$  denote the variables relating to hydrophobic, electronic and steric characteristics of the substituent, respectively;  $D$  is a dummy variable,  $a$ ,  $b$ ,  $c$  and  $d$  are coefficients of the variables, and  $e$  is the constant term;  $X_{ij}$  represents the presence of a particular substituent, and  $G_{ij}$  the contribution of that substituent to the potency. The term,  $\mu$ , has different implications depending on the model employed: it is the overall average of activities for the Free—Wilson model [3], but the calculated potency of the unsubstituted analog for the Fujita—

Ban model [6]. Equation (1) is the fundamental equation of the Hansch-Fujita method, whereas eqn. (2) represents its extended form.

Figure 1 depicts the main algorithm of QDD. The flow is controlled entirely by function-key (FK) and the concomitant alphanumeric-key (ANK) operations. Major roles of the FK operation are to: (a) display atomic coordinates; (b) display a molecular figure; (c) rotate a molecule; (d) rotate an atomic group; (e) combine substituents with a parent, and display the resulting derivative and its potency; (f) prepare hard copies; (g) display the result of trials. With regard to (a), (d), (e) and (g), other information is required from the ANK on the console. The role of (e) is the heart of the program, where substituents to be introduced into the parent are entered from the ANK, the desired derivatives are formed, and then the structure shows up on the CRT with its calculated potency.

In the QDD program, the original ANK input codes were devised and

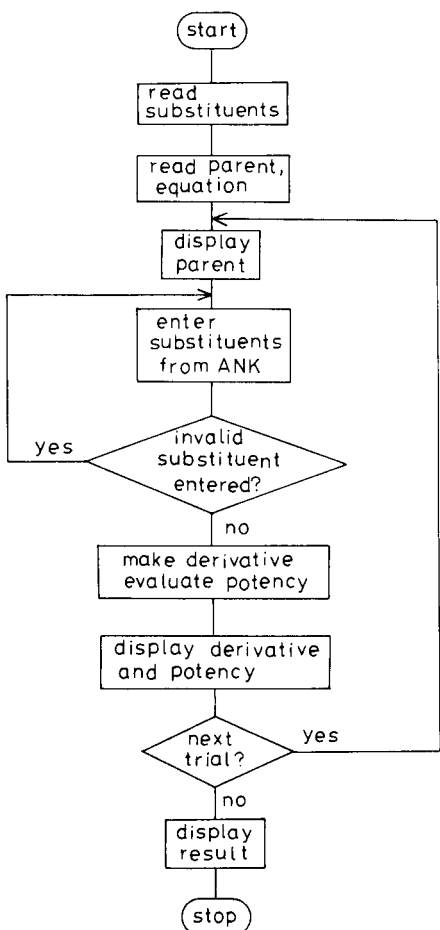


Fig. 1. Flow diagram of QDD.

TABLE 1

List of possible substituents (337 kinds) represented by the ANK input codes

No.	Formula	No.	Formula	No.	Formula	No.	Formula
1	BR	44	CH2BR	86	C2H4BR	128	CH=CHME
2	CL	45	CH2CL	87	CHCLME	129	CH2CH=CH <sup>f</sup>
3	F	46	CH2F	88	C2H4CL	130	CH2COME
4	SO2F	47	CH2I	89	NHCOME	131	COET
5	SF5	48	NHCHO	90	CONHME	132	SCOET
6	I	49	CONH2	91	CH2CONH2	133	OCOET
7	IO2	50	CH=NOH	92	CH=NOME	134	CO2ET
8	NO	51	OCONH2	93	NHCO2ME	135	CH2OCOM <sup>g</sup>
9	NO2	52	CONHOH	94	NHCSME	136	C2H4CO2H <sup>h</sup>
10	ONO2	53	CH2NO2	95	CSNHME	137	CHBRET
11	N3	54	ME	96	CH=NNHCSNH2	138	CHCLET
12	O-	55	HGME	97	ET	139	C3H6CL
13	SO2-	56	NHCONH2	98	OET	140	NHCOET
14	SO3-	57	NHCSNH2	99	CHOHME	141	C2H4CONH <sup>i</sup>
15	H	58	OME	100	CH2OME	142	CONHET
16	OH	59	CH2OH	101	C2H4OH	143	CH=NOET
17	SO3H	60	SOME	102	SOET	144	NHCO2ET
18	SH	61	SO2ME	103	CHOHCH2OH	145	NHCSET
19	ASO(OH)2	62	OSO2ME	104	CH2SO2ME	146	CHME2
20	B(OH)2	63	SME	105	SO2ET	147	C3H7
21	NH2	64	SEME	106	SET	148	NHCONHET <sup>j</sup>
22	NHOH	65	NHME	107	SEET	149	NHCSNHET <sup>k</sup>
23	SO2NH2	66	NHSO2ME	108	NHET	150	OCHME2
24	NH3+	67	NH2ME+	109	NME2	151	OC3H7
25	NHNH2	68	CH2NH3+	110	SO2NME2	152	COHME2
26	NHSO2NH2	69	C2F5	111	NHSO2ET	153	CHOMEME
27	CBR3	70	CCH	112	N(SO2ME)2	154	CHOHET
28	5-CL-1-TZL <sup>a</sup>	71	NHCOCF3	113	N3ME2	155	CH2OET
29	N=CCL2	72	OCF2CHF2	114	CH=NNHCO-HYD <sup>c</sup>	156	CH2CHOHM
30	CCL3	73	CH=CHCL	115	POME2	157	C2H4OME
31	CF3	74	CH2CN	116	PO(OME)2	158	SOC3H7
32	OCF3	75	CH=CHNO2	117	OPO(OME)2	159	SO2C3H7
33	SO2CF3	76	CH2SCN	118	PME2	160	SCHME2
34	SCF3	77	CH=CH2	119	SME2+	161	SC3H7
35	CN	78	NHCOCH2CL	120	CF(CF3)2	162	SEC3H7
36	NCO	79	COME	121	C(OH)(CF3)2	163	CHMENHMI
37	NCS	80	SCOME	122	CH=CHCN	164	CH2NME2
38	SCN	81	OCOME	123	CH2CCH	165	NHC3H7
39	CO2-	82	CO2ME	124	CH=CHCHO	166	NHSO2C3H7
40	NHCN	83	CH2CO2H	125	CH=CHCO2H	167	NME3+
41	1-TZL <sup>b</sup>	84	OCH2CO2H	126	C-C3H5 <sup>d</sup>	168	CH2NHME2
42	CHO	85	CHBRME	127	CME=CH2	169	PME3+
43	CO2H						

<sup>a</sup>5-Cl-1-tetrazolyl. <sup>b</sup>1-tetrazolyl. <sup>c</sup>CH=NNHCONH<sub>2</sub>. <sup>d</sup>C: cyclo. <sup>e</sup>1-pyrryl. <sup>f</sup>M: *meta*. <sup>g</sup>P: *para*. <sup>h</sup>ortho. <sup>i</sup>2-phenanthryl. <sup>j</sup>I-BU: isobutyl.

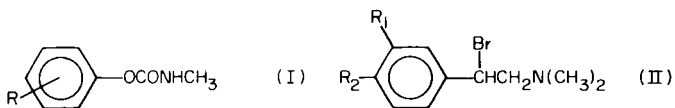
No.	Formula	No.	Formula	No.	Formula	No.	Formula
70	SIME3	212	PET2	254	SO2PH	296	CH2SPH
71	CH=C(CN)2	213	CH2NME3+	255	OSO2PH	297	C-C7H13
72	2-FURYL	214	CH2SIME3	256	NHPH	298	CH2-C-C6H11
73	3-FURYL	215	4-PYRIDYL	257	NHSO2PH	299	CME2CME3
74	2-THIENYL	216	CH=CHCOET	258	2,5-ME-PYRL <sup>h</sup>	300	CET3
75	3-THIENYL	217	CH=CHCO2ET	259	CH=CHCOC3H7	301	CHMECH2CME3
76	1-PYRL <sup>e</sup>	218	C-C5H9	260	CH=CHCO2C3H7	302	CH(C3H7)2
77	CH=CHCOME	219	CH=NOC4H9	261	C-C6H11	303	CH2SIET3
78	CH=CHCO2ME	220	NHCO2C4H9	262	O-B-GLU <sup>i</sup>	304	CCPH
79	C-C4H7	221	CHET2	263	CH2-A-PIP <sup>j</sup>	305	2-PHENANT <sup>n</sup>
80	CH2-C-C3H5	222	CH2CME3	264	CHMECME3	306	2-ANTHRYL
81	CH2CH=CHME	223	C2H4CHME2	265	C2H4CME3	307	3-INDOLYL
82	COC3H7	224	C5H11	266	C3H6CHME2	308	CH=CHPH
83	SCOC3H7	225	OCH2CME3	267	OCHMECME3	309	CH=NNHCOPH
84	OCOC3H7	226	OC2H4CHME2	268	CHMECOHMEET	310	N=CHPHOME-P
85	CH2CO2ET	227	OC5H11	269	CHOHC5H11	311	NHCOPHOME-P
86	CO2C3H7	228	CHMECOHME2	270	CHETCOHME2	312	CHMEPH
87	C3H6CO2H	229	CHOHCME3	271	PO(OC3H7)2	313	C2H4PH
88	NHCO3H7	230	CHOHC4H9	272	C3H6NME3+	314	N=CMENHPH
89	NHCOCHME2	231	CH2OCH2CHME2	273	PHCF3-M	315	C2H4OPH
90	CONHC3H7	232	CH2OC4H9	274	PHCF3-P	316	CH2SCH2PH
91	CH=NOC3H7	233	C3H6OET	275	PHCN-M	317	CME2OCH2CME3
92	NHCO2C3H7	234	C4H8OME	276	PHCN-P	318	CME2CH2OCME3
93	NHCSC3H7	235	CH(OET)2	277	2-BENZOXAZL <sup>k</sup>	319	PO(C4H9)2
94	CME3	236	CH(SET)2	278	2-BENZTHIAZL <sup>l</sup>	320	C2H4SIET3
95	CHMEET	237	CH2NET2	279	COPH	321	CH=CHCOPHNO2
96	CH2CHME2	238	CH2PO(OET)2	280	OCOPH	322	CH=CHCOPH
97	C4H9	239	C2H4NME3+	281	CO2PH	323	CHMECH2PH
98	OCH2CHME2	240	C6CL5	282	CH2PHCL-P	324	CHETPH
99	OC4H9	241	C6F5	283	CH2OPHCL-P	325	C3H6PH
00	COMEME2	242	PH(NO2)3	284	N=CHPH	326	CME2-C-C6H11
01	CHOETME	243	PHBR-M <sup>f</sup>	285	CH=NPH	327	CH(I-BU)2 <sup>o</sup>
02	CHOHC3H7	244	PHBR-P <sup>g</sup>	286	NHCOPH	328	1-NAPHTHYL
03	CH2COHME2	245	PHCL-M	287	PHME-O <sup>m</sup>	329	C4H8OPH
04	CH2CHOHET	246	PHCL-P	288	PHME-M	330	1-ADAMANTYL
05	C2H4OET	247	PHF-M	289	PHME-P	331	NPH2
06	CH2OC3H7	248	PHF-P	290	CH2PH	332	POPH2
07	NHC4H9	249	PHNO2-M	291	CHOHPH	333	PPH2
08	NET2	250	PHNO2-P	292	CH2OPH	334	CHPH2
09	P(OET)2	251	PH	293	CH2SO2PH	335	CH(OPH)2
10	PO(OET)2	252	N2PH	294	OSO2PHME-P	336	CO2CHPH2
11	OPO(OET)2	253	OPH	295	SO2PHOME-P	337	CMEPH2

2,5-dimethyl-1-pyrryl. <sup>i</sup>0-β-glucosyl. <sup>j</sup>CH<sub>2</sub>-α-piperidyl. <sup>k</sup>2-benzoxazolyl. <sup>l</sup>2-benzthiazolyl. <sup>m</sup>O:

employed to retrieve the substituent data. The input codes for substituents were selected in accordance with the following requirements, for easy inference of chemical structures: (a) expression by less than 12 characters; (b) use of capital letters instead of lower case, subscripts and superscripts, e.g.,  $\text{NH}_3^+ \rightarrow \text{NH3}^+$  or  $\text{Br} \rightarrow \text{BR}$ ; (c) adoption of obvious abbreviations, e.g., methyl  $\rightarrow$  ME, ethyl  $\rightarrow$  ET, phenyl  $\rightarrow$  PH; (d) use of chemical names or abbreviations for substituents difficult to describe by linear formulae, e.g., 3-thienyl  $\rightarrow$  3-THIENYL or *O*- $\beta$ -glucosyl  $\rightarrow$  O-B-GLU.

The data bank for substituents constitutes one of the important components of the program. At present, 337 kinds of substituents are present in the list which can be called on the CRT screen by the ANK operation. They are listed in Table 1 with the ANK input codes. For each substituent in Table 1, twenty different physicochemical parameter values [7, 8] and the Cartesian coordinates of constituent atoms are stored. Details of the physicochemical parameters are summarized in Table 2, where parameters 1 and 2 express the hydrophobic properties, parameters 3 and 4 indicate hydrogen-bonding properties, parameters 5–12 express steric properties, and parameters 13–20 electronic properties. The list of physicochemical parameter values for the 337 substituents can be printed out if necessary. The coordinates of constituent atoms of the substituent were generally evaluated for the presumed molecular model built up by using average atomic distances in the literature [9, 10]. However, bonds longer than 1.70 Å or shorter than 1.05 Å were adjusted to 1.70 Å or 1.05 Å, respectively, in connection with the limitations of the display program used (NAMOD). Thus, bonds such as C–Br, C–Cl and C–S, are estimated to be shorter than the actual values, and those such as O–H and N–H are estimated to be longer than the actual values.

Figure 2 gives an example of the frames displayed on the CRT, when the Hansch–Fujita method was applied with the QDD program, to the inhibitory activities of phenyl-*N*-methylcarbamates (I) against bovine erythrocyte acetylcholinesterase [11].



Biological potencies were evaluated by using the equation described elsewhere [8]. Figure 2(a) shows a perspective of the parent molecule I. The user next enters the input codes of substituents (*p*-mesyl) necessary to form a desired derivative from the ANK, and the molecular diagram is displayed in perspective with its calculated potency (Fig. 2b). These operations may be repeated by changing the substituents R1–R5 as often as necessary, and the list of substituents with the calculated potencies expressed as  $\log(1/C)$  are then displayed on the CRT.

Figure 3 shows some CRT frames for the Free–Wilson method applied to the adrenergic blocking potencies of *N,N*-dimethyl-2-bromophenethyl-



TABLE 2

Available physicochemical parameters<sup>a</sup>

No.	Symbol	Meaning	No.	Symbol	Meaning
1	$\pi$	Hansch—Fujita pi constant	13	$\sigma_m$	Hammett sigma constant
2	$f$	Nys—Rekker fragment constant	14	$\sigma_p$	
3	H-A	Hydrogen acceptor	15	$F$	Swain-Lupton constant
4	H-D	Hydrogen donor	16	$R$	
5	$MR$	Molar refractivity	17	$\sigma^*$	Taft inductive constant
6	$E_s$	Taft steric parameter	18	$E$	Frontier constant <sup>b</sup>
7	$E_s^c$	Hancock corrected steric parameter	19	$Re$	
8	$L$		20	$I$	
9	$B_1$				
10	$B_2$	STERIMOL parameter			
11	$B_3$				
12	$B_4$				

<sup>a</sup>Further details of these parameters except 18–20 are available [7]. <sup>b</sup>Details have been given [8].

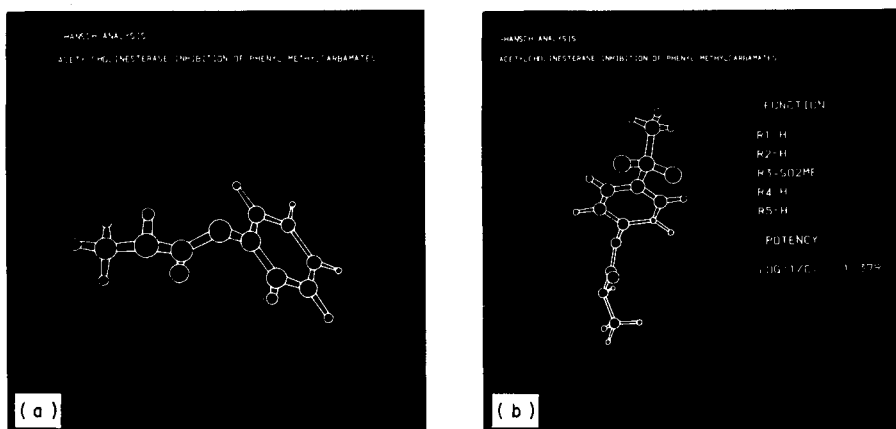


Fig. 2 The CRT frames for the Hansch—Fujita method applied to bovine erythrocyte acetylcholinesterase inhibition of phenyl-*N*-methylcarbamates: (a) parent molecule (phenyl-*N*-methylcarbamate); (b) *p*-mesylphenyl-*N*-methylcarbamate and its calculated potency.

amines (II) [12]. Biological potencies were evaluated by using the equation for the Fujita—Ban model. Again, the parent molecule (II) and a selected derivative with its calculated potency are displayed; the process is repeated and eventually a list of the R1 and R2 substituents with the calculated potencies of the derivatives can be displayed.

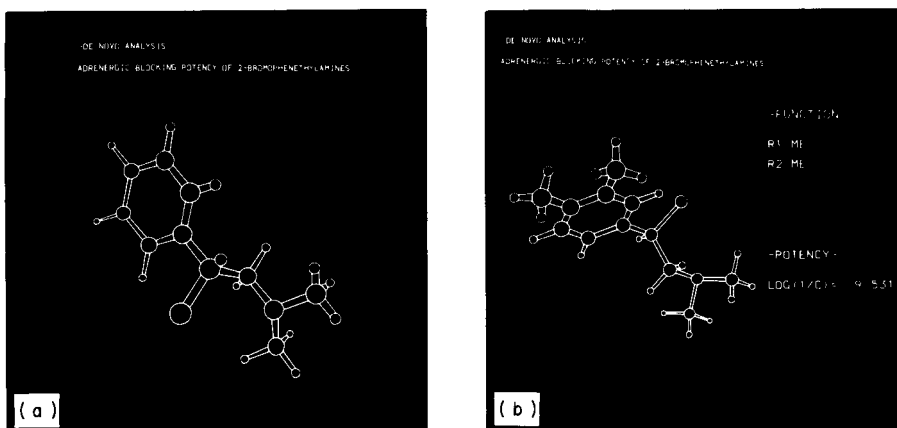


Fig. 3. The CRT frames for the Free—Wilson method on adrenergic blocking potencies of *N,N*-dimethyl-2-bromophenethylamines: (a) parent molecule (*N,N*-dimethyl-2-bromophenethylamine); (b) *N,N*-dimethyl-2-bromo-(*m,p*-dimethyl)phenethylamine and its calculated potency.

## DISCUSSION

When candidates are designed after the equation for estimation in the Hansch—Fujita and the Free—Wilson models have been established, they should be comprehensively evaluated by considering not only calculated potencies, but also the stability of the molecule, tractability, synthetic feasibility, availability or cost of raw reagents, etc. In particular, the steric properties of the molecule may give clues in inferring synthetic difficulty, molecular stability and sometimes the mechanism of drug action. In designing candidate molecules, it may be sensible to construct molecular models by hand in order to assess molecular conformation and steric problems, though this may be time-consuming if many derivatives are to be examined. The QDD program may permit the derivative to be visualized by simple FK and ANK operations, with prompt display of the resulting molecular model on the CRT together with its calculated potency. The molecular models displayed on the CRT provide very good perspectives (Figs. 2 and 3), and the user can easily change the direction of projection by simple molecular rotation. Conformational changes can be visualized by rotating a particular atomic group around a single bond axis if necessary, so that information similar to that obtained by three-dimensional models can be achieved. The program may permit, to some extent, prediction of the active site of the receptor and the mechanism of drug action, though the CRT sometimes displays confusing figures for large molecules.

The prophet system also employs computer graphics in research on structure—activity relationships [13–15] but no details have so far been reported on graphic programs such as QDD. The processing of optimizing

biological potencies by chemical modification forms one of the important stages in developing synthetic drugs and agricultural chemicals, in relation to the problem of increasingly resistant bacteria in chemotherapeutic agents and antibiotics. The Hansch—Fujita and Free—Wilson methods are expected to be more extensively applied in future, and the QDD program will improve their efficiency.

The author thanks Dr. Yoshitaka Beppu, Department of Physics, Faculty of Science, Nagoya University, for the source list of NAMOD and for his valuable advice on computer programming.

#### REFERENCES

- 1 E. C. Olson and R. E. Christoffersen (Eds.), *Computer-Assisted Drug Design*, ACS Symposium Series No. 112, American Chemical Society, Washington, DC, 1979.
- 2 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86 (1964) 1616.
- 3 S. M. Free and J. M. Wilson, *J. Med. Chem.*, 7 (1964) 398.
- 4 Y. Beppu, NAMOD, Program No. 370, Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, 1978.
- 5 J. J. Rhee and D. Blocher, COORD/1130, Program No. 226, Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, 1973.
- 6 T. Fujita and T. Ban, *J. Med. Chem.*, 14 (1971) 148.
- 7 C. Hansch and A. J. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979, pp. 65—168.
- 8 T. Esaki, *J. Pharm. Dyn.*, 3 (1980) 562.
- 9 J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory*, McGraw-Hill, New York, 1970, p. 111.
- 10 L. Pauling, *The Nature of the Chemical Bond*, 3rd edn., Cornell University Press, Ithaca, New York, 1960, p. 224.
- 11 T. Nishioka, T. Fujita, K. Kamoshita and M. Nakajima, *Pestic. Biochem. Physiol.*, 7 (1977) 107.
- 12 H. Kubinyi and O.-H. Kehrhahn, *J. Med. Chem.*, 19 (1976) 1040.
- 13 C. Hansch, private communication (1977).
- 14 P. H. Lenz, private communication (1977).
- 15 Y. C. Martin, private communication (1979).

## THE *ORTHO* EFFECT IN QUANTITATIVE STRUCTURE–ACTIVITY CORRELATIONS

TOSHIO FUJITA

*Department of Agricultural Chemistry, Kyoto University, Kyoto 606 (Japan)*

(Received 23rd January 1981)

### SUMMARY

A procedure for assessment of the effects of *o*-substituents in physical-organic systems by means of linear combination of substituent parameters is described and its applicability to various data sets on reactivity and equilibrium is shown. The procedure seems to provide a physicochemical background to the quantitative analysis of structure–activity relationships for biologically active aromatic *ortho* compounds. The *ortho* effect in biological activity is analyzed in terms of a linear combination of the electronic parameter  $\sigma_p$  and field-inductive  $F$ , hydrophobic  $\pi$  and steric  $E_s$  parameters. Depending on the type of steric effect, other parameters may be used. The model is shown to be applicable to several biological activity data sets of aromatic pesticide congeners including *o*-substituted derivatives.

Numerous quantitative structure–activity studies based on free energy-related physicochemical substituent parameters and regression analysis [1–6] have shown that the substituent effect on a certain biological activity of a series of congeneric compounds should be separable into electronic, hydrophobic, steric and other components. For the *m* and *p* substituted congeners, the Hammett  $\sigma$  constant can be used for the electronic effect of substituents [7].

The structure–activity relationship is analyzed quantitatively by means of

$$\log (1/C) = \rho \sigma + \sum e_i E_i + \text{constant} \quad (1)$$

where  $C$  is the concentration (or dose) of the congener giving a standard response (e.g.,  $EC_{50}$ ,  $LD_{50}$ ) on a molar basis;  $E_i$  corresponds to the free energy-related substituent parameters (e.g., hydrophobic and steric) other than those related to electronic effects;  $\rho$  and  $e_i$  are the regression coefficients of respective terms whose level of significance is examined statistically.

*Ortho*-substituted compounds, however, cannot be dealt with directly by eqn. (1) because the Hammett  $\sigma$  constant is usually not available for *o*-substituents. Generally, *o*-substituents exert various proximity effects such as steric and direct field electric effects which are otherwise insignificant, but these effects contribute to varying degrees in different systems, and the effect of *o*-substituents cannot be expressed by a single generally applicable set of “ $\sigma_o$ ” parameters in physical-organic systems. Thus, in order to develop

a model similar to eqn. (1) for structure—activity studies of *o*-substituted compounds, as well as to provide a physicochemical background for such a model, a suitable procedure for establishing *o*-substituent effects was required. The procedure developed is based on linear combination of substituent parameters and can be incorporated into a model for quantitative structure—activity analysis of *ortho* compounds [8]. The procedure is outlined below, and examples of this application to sets of aromatic pesticide congeners including *o*-substituted derivatives are given.

#### THE *ORTHO* EFFECT IN PHYSICAL-ORGANIC SYSTEMS

To establish the model, the following five assumptions were made [8]. First, the total effect of *o*-substituents, except for those capable of internal hydrogen bonding, is composed of the “ordinary” polar effect that can be considered in common with *m*- and *p*-substituents, and proximity effects (polar and steric) inherently associated with *o*-substituents. Second, the “ordinary” polar effects of *o*-substituents and *p*-substituents are the same, i.e.,  $\sigma_o \equiv \sigma_p$ , for any reaction type [9]. Third, the *o*-substituents are closer to the reaction center than the *p*-substituents. Thus, the definition  $\sigma_o \equiv \sigma_p$  may underestimate the electronic effect for *ortho* compounds. This underestimation, or the difference from the “ordinary” effect, is taken into account by the “proximity” polar effect which is represented by the Swain—Lupton  $F$  constants [10] as improved and extended by Hansch et al. [11]. Fourth, the steric effect of *o*-substituents can be expressed by the Taft  $E_s$  constants [12] as modified by Kutter and Hansch [13]. Fifth, when the secondary steric effect of the *o*-substituent (e.g., steric inhibition of resonance of the side-chain functional group) is significant as, for instance, in reactions where phenoxide formation is critical, the “ordinary” electronic effect of *o*-substituents is given by  $\sigma_p$  ( $= \sigma_o$ ) while the effects from *m*- and *p*-substituents are expressed by  $\sigma_{meta}^-$  and  $\sigma_{para}^-$ , respectively. (The composite of  $\sigma_o$  ( $= \sigma_p$ ),  $\sigma_m$  and  $\sigma_p^-$  is denoted as  $\sigma^\#$ .)

With these assumptions, the reactivity data of a series of *o*-substituted derivatives can be expressed by

$$\log k_o = \rho \sigma_o + \delta E_s^{ortho} + fF_o + c \quad (2)$$

where  $\delta$  and  $f$  are susceptibility constants and  $c$  is the intercept. If the  $\rho$  value of eqn. (2) does not differ from that of the correlation for the same reactivity of the corresponding *m*- and *p*-derivatives, the data for the series including *o*-, *m*- and *p*-substituted derivatives can be combined into eqn. (3).

$$\log k_{o, m, p} = \rho \sigma_{o, m, p} + \delta E_s^{ortho} + fF_o + c \quad (3)$$

Then, the proximity effects of *o*-substituents are considered to be well separated from the ordinary polar effect. The susceptibility constants and the intercept are determined by regression analysis.

The  $F$  values used in eqn. (3) are the improved values [11] (see above). The scale of the original set is not correct. The  $E_s$  values in eqn. (3) are not

the Taft  $E_s^0$  values defined for the aromatic *o*-substituents [12]. Charton [14] advanced evidence that the  $E_s^0$  value is linked rather with the polar substituent effect despite the original definition, while the  $E_s$  value determined for the aliphatic system is a real steric parameter [14]. Kutter and Hansch [13] found a relationship such as

$$E_s = -1.839 r_v (\text{av}) + 3.484 \quad (n = 6, r = 0.996, s = 0.132) \quad (4)$$

for symmetrical substituents such as H and  $\text{CX}_3$  (X = H, Me or halogen) where  $r_v (\text{av})$  is the average van der Waals radius [13]. The steric constants of heteroatom substituents were estimated by using eqn. (4) not only for other symmetrical substituents such as halogens but also for unsymmetrical substituents such as  $\text{NR}_2$ , OR,  $\text{NO}_2$  and phenyl [13]. Here the extended  $E_s$  values were used along with the original values for the alkyl substituents [12]. The reference substituent was shifted to H for simplicity so that  $E_s(\text{H}) = 0$ .

Numerous data sets were found [8] to obey eqn. (3): *o*-, *m*- and *p*-substituted derivatives for diverse reactions and equilibria including acid dissociation, hydrolysis and formation of esters and amides, electrophilic and nucleophilic substitution reactions, addition reactions and physicochemical properties. Some examples are shown in Table 1 along with the corresponding correlations for the *m*- and *p*-substituted compounds.

Depending on the conditions and type of reaction, not all the terms in eqn. (3) are statistically significant. For example, for dissociation (sets 1 and 2) or ion-pair formation (set 3) and  $\text{Ph}_2\text{CN}_2$ -esterification (sets 4 and 5) of substituted benzoic acids, one or other of the proximity effects becomes insignificant when the reaction medium is changed from hydroxylic solvents (water or ethanol) to aprotic solvents (DMSO, DMF and benzene). For the dissociation of phenoxyacetic acids and the  $\text{Ph}_2\text{CN}_2$ -esterification of phenylacetic acids, the proximity effects are totally unimportant (sets 6 and 7) which is attributable to the reaction site being distant from the benzene ring. When the polar effect is not significant in critical steps associated with the reaction mechanism, neither the  $\sigma$  nor the  $F$  term is required as observed for the acid-catalyzed hydrolysis of benzamides (set 8).

These examples indicate that the effect of *o*-substituents can be established with the use of  $\sigma_p$  for the ordinary electronic effect and the significance of the proximity effects can be examined by adding the  $E_s$  and  $F$  parameters. Thus, a model of the same form as eqn. (1) should apply to quantitative analysis of the biological activity of *o*-substituted compounds.

#### ANALYSIS OF THE BIOLOGICAL ACTIVITY OF *ORTHO*-SUBSTITUTED COMPOUNDS

In general, the *o*-substituent effect in biological activity can be analyzed in terms of a linear combination of substituent parameters as in the equation

$$\log (1/C) = a\pi + \rho\sigma + \delta E_s + fF + c \quad (5)$$

TABLE 1

Correlation of reactivity data<sup>a</sup> with eqn. (3):  $\log k = \rho\sigma + \delta E_s + fF + c$ 

$\rho$	$\delta$	$f$	$c$	$n$	$n_o$	$s$	$r^b$
1. Dissociation of ArCOOH in H <sub>2</sub> O at 25°C; log $K_A$ (M) [15]							
1.000			-4.197	24	0	0.021	0.998
(±0.029)			(±0.011)				
0.950	-0.392	1.469	-4.179	36	12	0.087	0.987
(±0.099)	(±0.050)	(±0.179)	(±0.039)				
2. Dissociation of ArCOOH in 50% DMSO at 25°C; log $K_A$ (M) [16]							
1.345			-5.838	6	0	0.033	0.998
(±0.106)			(±0.038)				
1.302		0.832	-5.800	11	5	0.113	0.987
(±0.248)		(±0.411)	(±0.089)				
3. Ion-pair formation of ArCOOH with (PhNH) <sub>2</sub> C=NH in benzene at 25°C; log $K$ (M <sup>-1</sup> ) [17]							
1.789			-0.613	22	0	0.141	0.979
(±0.174)			(±0.068)				
1.678	-0.470		-0.598	30	8	0.190	0.964
(±0.194)	(±0.161)		(±0.085)				
4. Esterification of ArCOOH with Ph <sub>2</sub> CN <sub>2</sub> in EtOH at 30°C; log $k$ (M <sup>-1</sup> min <sup>-1</sup> ) [18, 19]							
0.877			0.032	14	0	0.032	0.994
(±0.061)			(±0.020)				
0.878	-0.193	0.718	0.024 <sup>d</sup>	24	10	0.053	0.989
(±0.079)	(±0.033)	(±0.143)	(±0.029)				
5. Esterification of ArCOOH with Ph <sub>2</sub> CN <sub>2</sub> in DMF at 30°C; log $k$ (M <sup>-1</sup> min <sup>-1</sup> ) [19]							
1.557			-1.294	5	0	0.068	0.993
(±0.346)			(±0.129)				
1.774		0.359 <sup>c</sup>	-1.388	13	8	0.081	0.989
(±0.206)		(±0.252)	(±0.058)				
6. Esterification of ArCH <sub>2</sub> COOH with Ph <sub>2</sub> CN <sub>2</sub> in EtOH at 30°C; log $k$ (M <sup>-1</sup> min <sup>-1</sup> ) [20]							
0.368 <sup>o</sup>			0.067	10	0	0.021	0.985
(±0.053)			(±0.021)				
0.382 <sup>o</sup>			0.068	19	9	0.023	0.983
(±0.036)			(±0.013)				
7. Dissociation of ArOCH <sub>2</sub> COOH in H <sub>2</sub> O at 25°C; log $K_A$ (M) [21]							
0.313 <sup>o</sup>			-3.188	17	0	0.031	0.950
(±0.056)			(±0.023)				
0.317 <sup>o</sup>			-3.186	25	8	0.034	0.947
(±0.047)			(±0.019)				
8. Acid hydrolysis of ArCONH <sub>2</sub> in H <sub>2</sub> O (HCl) at 100°C; log $k$ (M <sup>-1</sup> min <sup>-1</sup> ) [22]							
0.085 <sup>d</sup>			-1.735	14	0	0.058	0.502
(±0.092)			(±0.041)				
	0.707		-1.733	23	9	0.129	0.965
	(±0.087)		(±0.066)				

<sup>a</sup>For each set, the correlation on the first line is for *m*- and *p*-substituted derivatives, whereas the correlation on the second line includes *o*-substituted derivatives. Unless otherwise noted, the values of  $\rho$ ,  $\delta$  and  $f$  are justified by *t*-tests at better than the 99.5% level of significance. The figures in parentheses are the 95% confidence intervals. <sup>b</sup>The number of data used in the correlation is denoted by  $n$ ;  $n_o$  denotes the number of data of *o*-substituted derivatives included in the correlation;  $s$  is the standard deviation and  $r$  is the multiple correlation coefficient. The unit of equilibrium or rate constant is shown in parentheses after  $K_A$  or  $k$ . <sup>c</sup>Justified at a level between 97.5 and 99.0%. <sup>d</sup>Justified at a level between 90 and 95%.





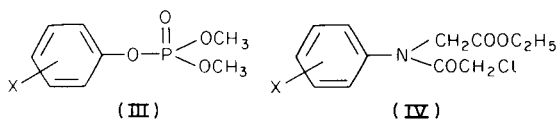
planthoppers and leafhoppers, which are major pests in rice production. The insecticidal activity of this class of compound (II) has been determined. The smaller brown planthoppers were liberated on rice seedlings which had been immersed in solutions containing various concentrations of the carbamates along with enough piperonyl butoxide to inhibit oxidative metabolic degradation. The concentration which caused 50% mortality after 24 h,  $LC_{50}$ , was estimated for each compound. For the *o*-substituted compounds, the equation established was  $\log(1/LC_{50}) = 0.953(\pm 0.428)\pi - 0.436(\pm 0.526)\sigma^0 + 0.723(\pm 0.375)E_s + 6.944(\pm 0.226)$ , for  $n = 13$  with  $s = 0.147$  and  $r = 0.871$ . In this equation, the  $\sigma^0$  term is significant only at the 90% level. The steric and hydrophobic effects are well separated.

Carbamate insecticides are believed to deactivate acetylcholinesterase in the nervous system of the insect [27]. The inhibitory activity against acetylcholinesterase prepared from planthoppers was therefore determined. The variation in the enzyme-inhibitor binding equilibrium constant ( $K_d^{-1}$ ) reflects the variation in the inhibitory activity. For the  $\log K_d^{-1}$  values of *o*-substituted derivatives, the equation derived was  $\log K_d^{-1} = 1.442(\pm 0.292)\pi + 0.645(\pm 0.398)\sigma^0 + 1.098(\pm 0.312)HB + 4.898(\pm 0.244)$  (with  $n = 14$ ,  $s = 0.194$  and  $r = 0.962$ ). Here  $HB$  is an indicator variable that takes a value of 1 for hydrogen-accepting substituents such as nitro, cyano and alkoxy groups, but is otherwise zero. The steric effect does not seem to be significant. The difference in the relative hydrogen-bonding effects of substituents between phases involved in binding with acetylcholinesterase and the 1-octanol/water partitioning phases used as the reference system to estimate  $\pi$  values is accounted for by the  $HB$  term [28].

The differences between the above equations for  $\log(1/LC_{50})$  and  $\log K_d^{-1}$  may reflect the factors other than oxidative degradation involved in biotransformation before the target is reached through the plant and/or insect body.

#### *Antiacetylcholinesterase activity of O,O-dimethyl o-substituted phenyl phosphates [29]*

This class of compounds (III) is generated in vivo by oxidative desulfuration of the corresponding phosphorothioates [30]. Among the phosphorothioates, the 3-methyl-4-nitro-, 3-chloro-4-nitro- and 4-cyano- derivatives have been widely used as selective insecticides with low mammalian toxicity. The insecticidal activity of the phosphorothioates is due to the inhibitory activity of the corresponding phosphates on acetylcholinesterase [27]. The kinetic constants were determined [29] for the inhibition reaction of this class of compound on an acetylcholinesterase preparation from house-fly heads. Similarly to the inhibition of acetylcholinesterase by phenyl *N*-methylcarbamates, the inhibitory potency is determined primarily by the binding step with the enzyme ( $1/K_d$ ). Because the number of *o*-substituted compounds showing appreciable enzyme inhibition is small, the  $\log(1/K_d)$  value of the *o*-compounds was analyzed together with that of the *m*-derivatives;  $\sigma_p^0$  and  $\sigma_m^0$  were used as the electronic parameters of the *o*- and *m*-substituents, respectively. The equation found was



$$\log(1/K_d) = 0.138\pi_{23}(\pm 0.053) + 2.246\sigma^0(\pm 0.205) + 2.884(\pm 0.115)$$

$$(n = 12, s = 0.046, r = 0.992) \quad (6)$$

In eqn. (6),  $\pi_{23}$  is the hydrophobic parameter for the *o*- and *m*-substituents. The correlation shown by eqn. (6) is very good and the standard deviation is low, indicating that no effect overlaps the ordinary electronic and hydrophobic effects which are equivalent for the *o*- and *m*-substituents. The *p*-substituents exert no hydrophobic effect on the binding; their electronic effect is correlated with  $\sigma^-$  by

$$\log(1/K_d) = 2.201\sigma^-(\pm 0.904) + 2.951(\pm 0.786) \quad (n = 7, s = 0.266, r = 0.942) \quad (7)$$

The  $\rho$  value of eqn. (7) is practically equal to that of eqn. (6). For *o*-, *m*- and *p*-substituted congeners,  $\sigma^\#$  can be used as a common electronic parameter to derive the equation

$$\log(1/K_d) = 0.176\pi_{23}(\pm 0.160) + 2.253\sigma^\#(\pm 0.304) + 2.892(\pm 0.215) \quad (n = 19, s = 0.153, r = 0.970) \quad (8)$$

which suggests that there is a through-resonance effect of *p*- but not *o*-substituents in the binding process.

#### *Herbicidal activity of N-chloroacetyl-N-phenylglycine esters [31]*

This class of compounds (IV) exhibits various degrees of herbicidal activity against annual grasses. To examine the selectivity in favor of the rice plant, the molar  $I_{50}$  concentrations which inhibit shoot elongation of a barnyard grass (*Echinochloa Cruss-galli* var. *frumentaceus*) and of rice plants to half the length of the control after 6 days were determined. The equations derived for the activity of *o*- and *m*-substituted derivatives, and some 2,6-disubstituted compounds, were as follows:

against the rice plant,

$$pI_{50} = -0.33\pi(\pm 0.18) - 0.95E_s^{ortho}(\pm 0.14) - 0.62\sigma(\pm 0.46) + 4.10(\pm 0.19) \quad (n = 28, s = 0.261, r = 0.959) \quad (9)$$

and against the barnyard grass,

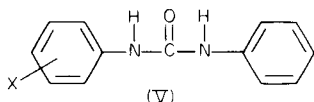
$$pI_{50} = -0.77E_s^{ortho}(\pm 0.16) - 0.22E_s^{meta}(\pm 0.20) + 3.99(\pm 0.24) \quad (n = 28, s = 0.295, r = 0.905) \quad (10)$$

The activity data of the *p*-derivatives, which were very low and did not vary significantly, were not included in the correlations. For the effects of 2,6-disubstitution,  $\Sigma\pi$ ,  $\Sigma\sigma$  and  $\Sigma E_s$  were used in the correlation. It is reasonable

to consider that the steric effect of *o*-substituents represented by the  $E_s^{ortho}$  term involves both intra- and inter-molecular types, while that of *m*-substituents expressed by the  $E_s^{meta}$  term is only for the intermolecular type. The negative sign of the  $E_s$  terms in eqns. (9) and (10) indicates that bulk is favorable to activity. Substituent effects providing selectivity for different plant species are evident from these two correlations, and information on selectivity between two plants can be obtained by comparing the equations.

#### Cytokinin-active substituted diphenylureas [32]

Bruce and Zwar [33] prepared a large number of substituted  $N,N'$ -diphenylureas exhibiting the cytokinin activity in varying degrees. Their activity data in terms of the minimum concentration ( $C$ ) to show detectable cell division on cultured tobacco pith were analyzed for compounds where



one of the benzene rings is unsubstituted (V) to give the equation

$$\log(1/C) = 0.90\sigma(\pm 0.33) - 0.85L_o(\pm 0.25) - 0.27L_p(\pm 0.22) + 1.04\pi_m(\pm 0.58) + 2.00(\pm 0.32) \quad (n = 39, s = 0.38, r = 0.91) \quad (11)$$

Not only monosubstituted but also polysubstituted derivatives were included in the correlation. For polysubstituted derivatives, the electronic effect is represented as  $\Sigma\sigma$ . The hydrophobic and steric effects of substituents at various positions in polysubstituted molecules are not additive but are primarily governed by substituents at particular positions. The  $L_o$  and  $L_p$  parameters are the length of the substituents at the *o*- and *p*-positions, respectively, along the "principal" bond axis as defined by Verloop et al. [34]. The  $p$  value, 0.90, suggests that an interaction with electron donors or nucleophiles at either the imino hydrogen or the carbonyl carbon is important for enhancing the activity. At the *o*- and *p*-positions, substituent length is unfavorable. At the *m*-position, very hydrophobic substituents enhance activity.

#### DISCUSSION

The above examples show that structure—activity correlations of biologically active *o*-substituted compounds can be obtained in essentially the same way as for the *m*- and *p*-substituted compounds. The "ordinary" as well as the position-specific and/or proximity effects of *o*-substituents can be analyzed by means of linear combination of physicochemical parameters. One of the most serious problems in dealing with this type of multiparameter correlation is collinearity among substituent parameters. This point was examined and it was confirmed that collinearity was not significant for most of the examples presented above.

Although further such examples can be expected, the procedure is not

applicable to all problems. For instance, there are several examples where two *o*-substituents together exert specific effects on biological activity. In the phenoxyacetic acid plant growth regulators, a 2-chloro substituent greatly enhances the activity of the 4-chloro compound (leading to 2,4-D) whereas the 2,6-dichloro analog shows very little activity [35]. In a study of the activity of *N*-benzoyl-*N'*-(4-chlorophenyl)urea derivatives against the larvae of the rice stem borer, a 2-fluoro substituent at the benzoyl moiety significantly lowers the activity of the unsubstituted benzoyl compound whereas the 2,6-difluoro substitution considerably increases the activity [36]. The effect of two *o*-substituents is apparently not additive in either example. The present procedure, which is primarily intended for assessment of the effect of single *o*-substituents and assumes the additivity of component effects, should be extended and modified to cover even those complex problems.

Finally, it should be emphasized that the analysis is just the first step in understanding the physicochemical mode of action of biologically active compounds. Numerous efforts will be needed from various directions to clarify the physicochemical significance of substituent effects at the (sub)-molecular level for biologically active compounds and for the receptors.

#### REFERENCES

- 1 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86 (1964) 1616.
- 2 C. Hansch, *Acc. Chem. Res.*, 2 (1969) 232.
- 3 C. Hansch, *J. Med. Chem.*, 19 (1976) 1.
- 4 T. Fujita, *Adv. Chem. Ser.*, 114 (1973) 1.
- 5 Y. C. Martin, *Quantitative Drug Design*, Dekker, New York, 1978.
- 6 S. H. Unger, *Drug Design*, 9 (1980) 48.
- 7 L. P. Hammett, *Physical Organic Chemistry*, 2nd edn., McGraw-Hill, New York, 1970, p. 347.
- 8 T. Fujita and T. Nishioka, *Progr. Phys. Org. Chem.*, 12 (1976) 49.
- 9 O. Exner, in N. B. Chapman and J. Shorter (Eds.), *Advances in Linear Free Energy Relationships*, Plenum, London, 1972, p. 27.
- 10 C. G. Swain and E. C. Lupton, *J. Am. Chem. Soc.*, 90 (1968) 4328.
- 11 C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikaitani and E. Lien, *J. Med. Chem.*, 16 (1973) 1207.
- 12 R. W. Taft, in M. S. Newman (Ed.), *Steric Effects in Organic Chemistry*, Wiley, New York, 1956, p. 556.
- 13 E. Kutter and C. Hansch, *J. Med. Chem.*, 12 (1969) 647.
- 14 M. Charton, *J. Am. Chem. Soc.*, 91 (1969) 615.
- 15 A. Albert and E. P. Serjeant, *Ionization Constants of Acids and Bases*, Methuen, London, 1962, p. 121.
- 16 M. Hojo, M. Utaka and Z. Yoshida, *Kogyokagaku Zasshi*, 69 (1966) 885.
- 17 M. M. Davis and H. B. Hetzer, *J. Res. Natl. Bur. Stand.*, 60 (1958) 569.
- 18 N. B. Chapman, J. Shorter and J. H. P. Utley, *J. Chem. Soc.*, (1962) 1824.
- 19 A. Buckley, N. B. Chapman, M. R. J. Dack, J. Shorter and H. M. Wall, *J. Chem. Soc. B*, (1968) 631.
- 20 N. B. Chapman, J. R. Lee and J. Shorter, *J. Chem. Soc. B*, (1969) 769.
- 21 N. V. Hayes and G. E. K. Branch, *J. Am. Chem. Soc.*, 65 (1943) 1555.
- 22 E. Reid, *Am. Chem. J.*, 24 (1900) 397.
- 23 T. Fujita, J. Iwasa and C. Hansch, *J. Am. Chem. Soc.*, 86 (1964) 5175.

- 24 S. Nakagawa, K. Nishimura, T. Fujita and M. Nakajima, unpublished work.
- 25 A. Bondi, *J. Phys. Chem.*, 68 (1964) 441.
- 26 K. Kamoshita, I. Ohno, K. Kasamatsu, T. Fujita and M. Nakajima, *Pestic. Biochem. Physiol.*, 11 (1979) 104.
- 27 W. N. Aldridge and E. Reiner, *Enzyme Inhibitors as Substrates*, North-Holland, Amsterdam, 1972.
- 28 T. Fujita, T. Nishioka and M. Nakajima, *J. Med. Chem.*, 20 (1977) 1071.
- 29 K. Kamoshita and T. Fujita, unpublished work.
- 30 J. C. Gage, *Biochem. J.*, 54 (1953) 426.
- 31 A. Fujinami, T. Satomi, A. Mine and T. Fujita, *Pestic. Biochem. Physiol.*, 6 (1976) 287.
- 32 H. Iwamura, T. Fujita, S. Koyama, K. Koshimizu and Z. Kumazawa, *Phytochemistry*, 19 (1980) 1309.
- 33 M. I. Bruce and J. A. Zwar, *Proc. Roy. Soc. London, Ser. B*, 165 (1966) 245.
- 34 A. Verloop, W. Hoogenstraaten and J. Tipker, *Drug Design*, 7 (1976) 165.
- 35 R. M. Muir and C. Hansch, *Plant Physiol.*, 28 (1953) 218.
- 36 Y. Nakagawa, K. Kitahara, T. Fujita and M. Nakajima, unpublished work.

## SELF-ADAPTING COMPUTER PROGRAM SYSTEM FOR DESIGNING ORGANIC SYNTHESSES

Z. HIPPE

*The I. Łukasiewicz Technical University, 35-959 Rzeszów (Poland)*

(Received 23rd January 1981)

### SUMMARY

The computer program SCANSYNTH is designed to assist in planning organic syntheses. The system features a new mode of data input/output, and is self-adapting, with simultaneous generation of possible synthesis schemes in the forward and reverse directions. Operational options include modeling of individual reactions, one-step reactions and multi-step reactions.

A promising area of chemical information systems is computer-aided forecasting of organic reactions and syntheses. There are two distinct approaches to the solution of such problems. The first, well represented by Ugi's group [1], is based on a mathematical model of constitutional chemistry which enables even novel, unknown reactions to be generated theoretically. In the second, but earlier, approach which was originated by Corey [2] and developed by Gelernter [3], Wipke and Gund [4] and others [5–10], information about known reactions is stored in the computer memory and used to generate a series of possible synthetic routes to a given complex molecule.

In the present paper, experience with the SCANSYNTH system [11–14] is outlined and general questions related to computer-aided assessment of organic synthetic routes, are briefly discussed. To provide a proper background for further discussion, the unusual features of SCANSYNTH compared with similar systems such as CYCLOPS, LHASA, SECS or SYNCHEM2, are first summarized.

### DESCRIPTION OF THE SYSTEM

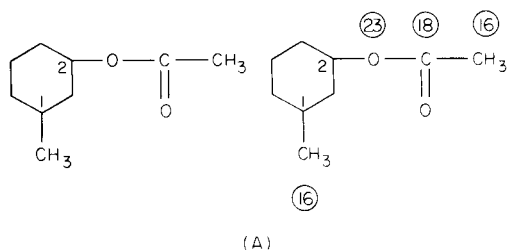
The main features of SCANSYNTH are: (i) the technique used for data input, (ii) its self-adaptability, (iii) the simultaneous forward and reverse generation of synthesis schemes, and (iv) three independent operating modes.

#### *Data input*

The data input and internal representation of the target molecule as well as of all the intermediates generated are important. The fragmentation code

CONOL-II [15] is used. The notation, very easy for coding and decoding, is applied in all the higher-level computer program systems developed in Poland. From the beginning of the project, the extremely awkward Wiswesser line-formula notation was avoided.

Within the machine, the molecular structure is stored and processed as a combined matrix-vector representation (Table 1). All connections between functional groups and ring systems (i.e., any single ring and/or set of condensed or spirane rings) are described by a functone connectivity matrix, FCM, whereas additional information on rings is stored in vectors. This method of representing the chemical structures within the computer shows two important features: it offers fast processing and has low memory requirements. For instance, molecule A containing 27 atoms may be described by means of a quadratic matrix having only  $5 \times 5 = 25$  elements, instead of 729.



(Structural formula)

CONOL-II code: 1C1-C2-2C1-C2-C2-C2-1C1, 1-C3, 2-0-3C-C3, 3 = 0.  
 Numerical code: 123 17 24 17 223 17 24 17 24 17 24 17 123 52  
 117 25 52 217 26 17 322 17 25 52 318 26 52

In Table 1, the connectivity matrix provides the information that functone no. 1 (ring system) is connected with a methyl group (no. 16) in node 1 by a single bond, whereas in the case of node 2, another functone ( $\text{—O—}$ , no. 23) is

TABLE 1

Internal representation of structure A.  
 (For explanation, see text)

Functone connectivity matrix, FCM<sup>a</sup>

	1	16	16	18	23
1	0	101	0	0	102
16		0	0	0	0
16			0	1	0
18				0	1
23					0

Ring system vector, RSV

6	1	0	0	0	16	0	23	0	0	0	1	23	17	24	17	2	23	17	24	17	24	17	24	17	24	17	1	23	0	16
1	2	3			4						5																		6	7

<sup>a</sup>Here, 1 = single bond, 2 = double bond, 3 = triple bond. 101 means that functone no. 16 is single-bonded to atom no. 1 of the ring system.

single-bonded. Similarly, functones 18 and 23 (C=O and —O—, respectively) as well as a second CH<sub>3</sub>— group and >C=O (16 and 18) are connected by single bonds. The ring system vector, RSV, carries all necessary information about the ring: six-membered (1), alicyclic (2), without heteroatoms (3), two nodes of the ring (first and third atoms) are connected with functones 16 and 23, respectively (4). Other elements of the vector show the ring notation in machine codes (5) or common atoms with other rings (6) whereas the last digit (7) indicates the number of elements used to describe the ring. [Other vectors are also applied here, to fix the number of rings in a ring system (in the present example, 1) to store the number of elements of the vector RSV (29) and to indicate the number of rings in the molecule analyzed.] This mixed representation can be easily reconverted to numerical code and again into the CONOL-II notation on output. Moreover, information stored in the data base of reactants and in the data base of reactions is in the form of a functone connectivity matrix, which improves the process of retrieval of data required for the system operation. The system is well endowed with versatile and fast subroutines for checking the correctness of the code or for rejecting unstable structures from a synthesis design.

### Self-adaptability

The idea of self-adaptability of the system was created during the work on algorithms for seeking a position (strictly speaking, the bond) in the molecule for the most effective reduction of its complexity. This bond is the point of origin of the process of generating the synthesis design on every level (Fig. 1).

A distinct problem was the field of application of rules for establishing the vital bond. It seemed that development of a general problem-solving procedure for use by the computer in establishing this bond is considered too difficult for a final solution to be achieved in the foreseeable future [16]. It was therefore decided to equip SCANSYNTH with partial problem-solving procedures, suitable for reverse simplification of specified types of structures, e.g., symmetric, polycyclic or chain structures. It was found, however, that the system was often useless even for simple structures, i.e. the essential bond was not found and no suggestion was given as to which chemical transformation should be selected for a particular step in the synthesis. For this reason, the most recent version of SCANSYNTH is endowed with a self-adapting function for automated modification of the generated structure. When a given procedure fails, the analyzed structure is modified appropriately.

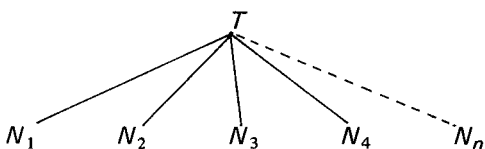
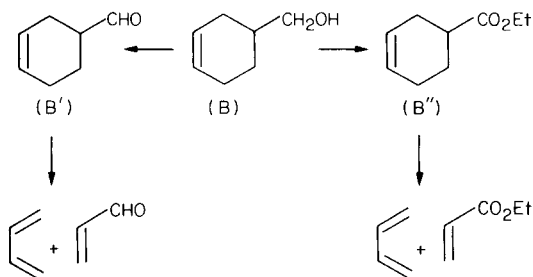


Fig. 1. First level of the synthesis design for target molecule *T*. Here,  $N \in \langle N_1, \dots, N_n \rangle$  is the set of possible intermediates of the first order (first generation).



This situation is illustrated by structure B, which must first be converted to structure B' or B'' and then a Diels–Alder reverse simplification may be successfully applied:



Another very powerful self-adapting procedure is connected with simplification of chain structures. In the forward direction, this procedure consists of forming a carbon–carbon bond, usually in the central part of a molecule, so that the main skeleton is built up. The basic operation of the algorithm relies on seeking the central C–C bond and defining activating chemical groups (e.g.,  $>\text{C}=\text{O}$ ,  $-\text{NO}_2$ ,  $\text{O}=\text{CH}$ ,  $\text{O}=\text{C}-\text{O}-\text{C}-$ ) adjacent to it. Altogether, forty-one types of C–C bonds (with various combinations of activating groups) as well as ten types of central fragments have been chosen. When an appropriate central bond cannot be found, a conversion graph (Fig. 2) is

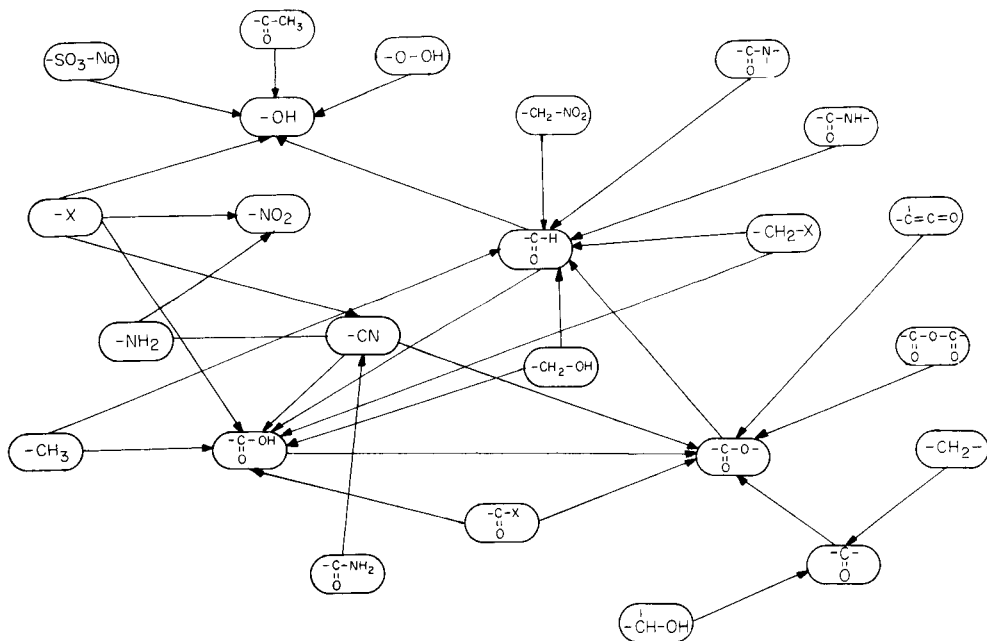
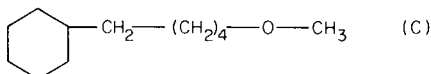


Fig. 2. Conversion graph for the self-adapting modification of chemical intermediate structure.

automatically applied to check if a desirable group connected with the selected central fragment may be preferred to an alternative group, obeying the rule of the lesser structural change.

For instance, in molecule C, no vital bond was found, but the central

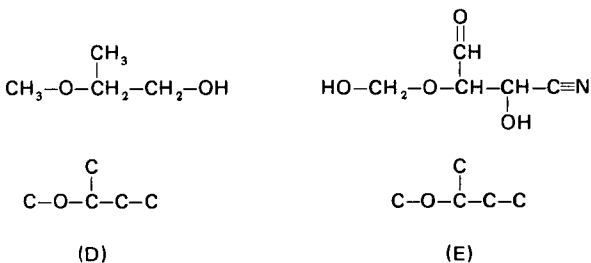


benzyl fragment was identified. After self-adapting conversion of  $-\text{CH}_2-$  to  $>\text{C}=\text{O}$ , the modified structure shows the important bond between the ring and the carbonyl group.

### *Simultaneous generation of the synthesis design in the reverse and forward directions*

An unusual feature of the SCANSYNTH system arises from the principal algorithm which was originally developed to reduce the machine time for discovery of the synthetic pathway. Essentially, the system itself looks for structural differences (or structural similarities) between the generated chemical intermediate and one or several elements of the set of reactants, stored in the computer memory. The algorithm which does this also indicates the sequence of chemical conversions necessary for rearrangement of the given reactant to yield the intermediate generated. The criteria of structural similarity are as follows. Structure R (reactant) is considered to be similar to structure S (chemical intermediate) provided that: (a) the skeleton of S contains the skeleton of R; (b) the remaining atoms (C, O, N, S and halogens) which differentiate the two skeletons belong to terminal functions in structure S; and (3) substituents in R have their counterparts in S, or are mutually interchangeable.

It should be explained that the skeleton is defined as the fragment of the molecule after elimination of so-called non-carbon terminal functions and hydrogens, and after replacement of multiple bonds by single bonds. For instance, the structures, D and E,



are similar, because the two skeletons are consistent and terminal functions can easily be reconverted (see conversion graph, Fig. 2).

This strategy has two important results. First, the synthetic pathway, on a given level of the synthesis design may begin from the reactant which is structurally very similar to the intermediate sought. This is, of course, a version of the rule used intuitively by chemists that the main part of the synthesized molecule is transferred unchanged from the starting material during the synthesis. Secondly, the synthetic pathway is first generated in reverse, but from a given level can also be generated in a forward direction. It is believed that in this manner many ineffective program loops are avoided, and so machine time is substantially decreased.

### *Systems options*

The executor of the SCANSYNTH system makes it possible to choose one of three allowed modes of operation: option R (reaction), option S (step), or option T (tree).

Option R is closely connected with the status of the data base of reactions. The data base is set up in open form, i.e., well equipped for automated completion, extension, truncation and modification of records. It is hoped that after suitable extension of the quantity of entries, the level of entropy of information will be higher than that available from even experienced chemists. However, apart from the number of reactions recorded, better exploitation of stored information is also needed. For this purpose, the system is equipped with an option for answering the general question: is the reaction  $X + Y \rightarrow Z$  feasible? Obviously, this question must be stated properly: the input data include information on the structures of product Z and reactant X (or X + Y). The computer recognizes substructure units in Z and X (possibly also in Y), searches for a model reaction from the data base, checks whether or not some structural units in the reactant will disturb the course of the reaction, and finally, prints the simulated reaction. This option is particularly helpful for industrial people and also in teaching at basic and advanced levels.

In option S, the system produces only the first generation of intermediates for a given target molecule. The one-step generation of intermediates plays a useful role in aiding conceptual work, particularly in studies of potential uses of industrial byproducts. Finally, by means of option T, full synthesis design is generated and different possible routes for the synthesis of a complex molecule are produced. The SCANSYNTH system appears to offer considerable flexibility in assisting the design of syntheses.

### REFERENCES

- 1 I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, G. Jochum and W. Schubert, *Angew. Chem.*, 91 (1979) 99.
- 2 E. J. Corey, *Q. Rev. Chem. Soc.*, 25 (1971) 455.
- 3 H. Gelernter, N. S. Sridharan, H. J. Hart, S. C. Yen, F. W. Fowler and H. J. Shue, *Top. Curr. Chem.*, 41 (1973) 113.
- 4 W. T. Wipke and P. Gund, *J. Am. Chem. Soc.*, 98 (1976) 8107.
- 5 R. Barone, M. Chanon and J. Metzger, *Rev. Inst. Fr. Pet.*, 28 (1973) 771.

- 6 M. Bersohn and A. Esack, *Chem. Rev.*, 76 (1976) 269.
- 7 P. E. Blower, Jr. and H. W. Whitlock, Jr., *J. Am. Chem. Soc.*, 98 (1976) 1499.
- 8 A. Weise, *Z. Chem.*, 15 (1975) 333.
- 9 Z. Hippe, *Proc. All-Union Conf. on Application of Computers in the Spectroscopy of Molecules*, Sib. Div. Acad. Sci. U.S.S.R., Novosibirsk, 1977.
- 10 R. Hippe and Z. Hippe, *Proc. IV Int. Conf. on Computers in Chemical Research and Education*, Sib. Div. Acad. Sci. U.S.S.R., Novosibirsk, 1978.
- 11 Z. Hippe, R. Hippe, M. Dec and W. Szumiło, *Proc. I Int. Conf. Software Reliability*, Książ, Techn. Univ. Press, Wrocław, 1979.
- 12 R. Hippe and Z. Hippe, *Proc. IV Int. Symp. System Modelling Control*, Polish Cybern. Soc., Zakopane, Poland, 1979.
- 13 Z. Hippe, R. Hippe, O. Achmatowicz, Jr., M. Dec, J. Lipiński, G. Kruczek, Z. Zielińska, E. Jędrzejec and W. Szumiło, *Proc. All-Union Conf. on Applications of Computers in Chemistry*, Sib. Div. Acad. Sci. U.S.S.R., Novosibirsk, 1980.
- 14 J. Gasteiger and Z. Hippe, in preparation.
- 15 Z. Hippe, R. Hippe, M. Dec and G. Kruczek, *Coding of Chemical Structures in the CONOL-II Notation*, Techn. Univ. Press, Rzeszów.
- 16 A. J. Thakkar, *Fortschr. Chem. Forsch.*, 39 (1973) 1.

## AUTOMATIC DEDUCTIVE SYSTEMS FOR CHEMISTRY

P. A. D. deMAINE\*

*Computer Science Department, The Pennsylvania State University, University Park,  
PA 16802 (U.S.A.)*

(Received 23rd January 1981)

### SUMMARY

Deductive systems are general programs intended to assist in the design of experiments and in data processing. In such systems no mathematical assumptions are made; only the fundamental law of conservation of energy/mass is used; and the systems have a predictive capability that does not use "state-of-the-art libraries". Two kinds of deductive systems are available: numeric deductive systems which have computative as well as predictive capabilities, and alphanumeric deductive systems which have only predictive capabilities and can be used to predict possible reaction paths. The current status of both kinds is reviewed, and examples of the use of a fully operational numerical deductive system are discussed.

Deductive systems solve class problems of interest to a significant number of scientific workers. In such systems no mathematical assumptions are made; only the fundamental law of conservation of mass and energy is used; and there are predictive facilities that do not use "state-of-the-art" libraries. Ideally, the implemented forms of such systems should be easy to use by workers with minimal computer skills. Within the framework of the available computer facilities there should be no restriction on either the size or complexity of the class problem that is to be solved. And no attempt should be made to impose standards with respect to the use of either mathematical methods or data-processing techniques.

So far, two kinds of chemical deductive systems have been identified. First, numeric deductive systems have computative capabilities in addition to the predictive facility that is used to determine those combinations of parameters for which data are needed to compute values for any subset of parameters. Second, alphanumeric deductive systems have only predictive capabilities and are used to predict those entities (*viz.* reactants) that may be combined in experiments to yield designated products.

The description and use of fully operational deductive systems will be found in a series of papers on automatic deductive systems, that are now being prepared for publication. In this paper, the current status of the entire

---

\*On leave to the Ballistic Missile Defence Advanced Technology Center, Research Park, Huntsville, Alabama 35807, U.S.A.

deductive system package is first discussed, and then some examples of the use of the fully operational CRAMS deductive system are given. This paper is intended to serve as an introduction to the series; automatic deductive systems.

### *Overview of proposed Deductive Systems Package*

So far nine interdependent systems have been indentified as intrinsic parts of the partially implemented Deductive Systems Package (Fig. 1); a system is defined here as a program with 10,000 or more statements. The status of each of those component systems is indicated in Table 1. The component systems are designed so that they can also be used on a stand-alone basis. It should be noted that there are prototypes for seven of the nine systems and that three of those (CURFIT, CRAMS and INTEGRAL) are available for production use.

When completed, the *transportable programming language* (TPL) system can be used to code machine- and configuration-independent programs (or systems) that will correctly compile and execute on virtually any medium (viz. PDP 11/45) to large (viz. CDC 7600 or IBM 3033) machine. Brief [1] and detailed [2, 3] descriptions of TPL are available. The TPL library has been described [4]. Here it is sufficient to note that TPL can be implemented for any high-level language (FORTRAN, COBOL, etc.) and that the minimal machine requirements to compile and execute any TPL program, no matter what its size, are a commercial high-level language compiler, at least 5000 sixteen-bit words for executing the compiled program, sufficient direct access storage for storing data and libraries, and I/O for communicating with the machine.

The TPL package consists of: (i) the TPL compiler that is coded in TPL and is therefore transportable, i.e., no recoding is necessary; (ii) the TPL

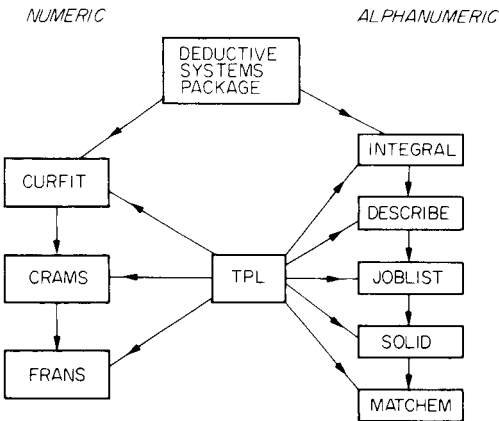


Fig. 1. Relationships among the interdependent component systems of the Deductive Systems Package. TPL will be eventually used to code all systems. CURFIT is a subsystem of CRAMS, and both CURFIT and CRAMS will be subsystems of FRANS. INTEGRAL, DESCRIBE, JOBLIST, and SOLID comprise the Retrieval Package. The functions of all systems are discussed in the text.

TABLE 1

Current status of the component systems for the Deductive Systems Package. D, I and L specify the status of design and implementation and the coding language used respectively.<sup>a</sup>

System	Status	System	Status
TPL (i) Language	D(c), I(-), L(-)	FRANS	D(a)
(ii) Library	D(c), I(pv), L(f/a)	INTEGRAL	D(c), I(pv), L(a)
(iii) Compiler	D(a), I(pr), L(f)	DESCRIBE	D(c)
(iv) SMSP	D(s)	JOBLIST	D(c), I(pr), L(f)
CURFIT	D(c), I(pv), L(f/a)	SOLID	D(c), I(pr), L(a)
CRAMS	D(c), I(pv), L(f)	MATCHEM <sup>b</sup>	D(a), I(pr), L(pl/1)

<sup>a</sup>The qualifiers in parentheses mean: for D: c, complete; s, started; a, advanced; for I: pr, prototype; pv, production version; for L: a, IBM Basic Assembler Language; f, FORTRAN; tpl, TPL; pl/1, PL/1; f/a, predominantly FORTRAN with some IBM BAL. Thus D(c), I(pv) and L(f/a) mean that the design of CURFIT is complete; there is a production version; and it is coded predominantly in FORTRAN with some assembler-coded routines.

<sup>b</sup>Prototypes are the CICLOPS and EROS implementations of the model by Ugi and co-workers (see below).

library, also coded in TPL, that supports replacements for non-transportable instructions in the parent high-level language, and instructions not available in any high-level language; (iii) the specifications for the TPL language, the syntactic and semantic specifications of which are sufficiently detailed to justify the claims that TPL is a new language; and (iv) the so-called small machine support package, SMSP, that is needed if programs that normally execute on large machines like the IBM 3033 are to execute on a medium machine like the PDP 11/45.

It can be demonstrated that the efficiency of execution of a program coded in TPL is ultimately determined by the efficiency of the commercial compiler, i.e., a program coded in TPL will not be less effective than a similar one coded in the parent high-level language.

It is planned ultimately to recode all our systems in TPL so that they will successfully compile and execute on virtually any medium to large machine. At this time the production versions of CURFIT and CRAMS are coded in IBM FORTRAN IV; and the production version of INTEGRAL is coded in IBM BAL. A brief discussion of components for the two kinds of deductive systems follows.

#### ALPHANUMERIC DEDUCTIVE SYSTEMS

This part is directed towards realizing an operational, general implementation of the mathematical model of Ugi and co-workers for synthesis of compounds [5-8] that can be easily used by workers with minimal computer skills. In the Ugi model, information about the electronic configurations of all reactants and all products is recorded in the so-called beginning, B, and end, E, matrices, and they are related via the reaction, R, matrix thus:  $B + R = E$ . The commutative rules apply and R can be replaced by  $\Sigma_i R_i$ .

The preceding equation can then be rewritten:

$$B + R = B_1 + \sum_{i=2} R_i = B_2 + \sum_{i=3} R_i = \dots = E.$$

The  $B_1, B_2, \dots$  are identified as intermediate compounds and thus the corresponding  $R_i$  are concise descriptions of the general chemical rules for transforming one intermediate (or reactant) into another. The feasibility of a particular reaction is determined by two factors: (i) every  $R_i$  must be a valid description of the chemical rules; (ii) the intermediate that is formed in every intermediate reaction must exist.

The three major components that are needed to implement successfully a general software version of the Ugi et al. model that will be both efficient and easy to use are as follows. First, a generalized data structure is needed for describing and manipulating the B, E and R matrices, that can also be used to retrieve and store information about parts or entire intermediates. This data structure must support all the normal arithmetic and matrix operations and, in addition, support the several special operations, e.g., rotating or counting certain characteristics, needed to manipulate representations of chemical species. Second, the user interface must support a chemistry-like language that is familiar to researchers and be both self-correcting (if possible) and self-educating. Third, there must be a high-speed retrieval system that has the following characteristics: (i) fully automatic with respect to both the allocation of computer resources and the processing of all queries; (ii) efficient with respect to the use of both storage and communications resources; (iii) independent of both information and question-type, i.e., the processing of any query must not depend on the form of the query (explicit or implicit); (iv) the search times should be both small and bounded, i.e., the maximum amount of CPU time needed to process any query should be independent of the amount or type of information in the system; (v) there must be a security system that can be easily used to prevent unauthorized access to any item of information.

The four support systems for the alphanumeric deductive systems package (INTEGRAL, DESCRIBE, JOBLIST and SOLID) together fully meet the general specifications outlined above. These four systems are also intended to serve, either singly or together, as the nucleus for an economically viable distributed information network [1]. The five component systems for the planned alphanumeric deductive system are discussed next.

INTEGRAL is a reversible compressor that compresses/decompresses digitized data at high speeds without loss of even a single binary bit of significant information. Details of the fully operational production version, which is coded in IBM Basic Assembler, are available [9, 10]. General descriptions of the techniques used have been given [11–13]. Here it is sufficient to note that the current version [9] yields savings as high as 99.0%, with a median between 60% and 80%, at decompression speeds as high as 470,000 bytes/second (= 3,760,000 BAUD) on the IBM 360/67(1). With the IBM 370/168 the savings are the same but compression and decompression speeds have exceeded 3,200,000 and 8,000,000 BAUD respectively.



A hard-wired version should operate at speeds above 8,000,000 BAUD. It is planned to recode INTEGRAL in TPL when that system is completed.

The DESCRIBE data-description language [14] is designed to describe chemical information and data in easily understood form. The DESCRIBE system, which is still to be implemented, will convert information coded in the DESCRIBE language to the JOBLIST form (see below).

The general data structure that is used is called JOBLIST [15, 16]. The partially implemented FORTRAN-coded JOBLIST system creates, maintains and manipulates the JOBLIST data structure. The relationships between DESCRIBE, JOBLIST and SOLID [15] are as follows. The DESCRIBE language is used to describe both information and the operations that are to be performed. The JOBLIST system constructs the internally used data form, called JOBLIST, and executes all arithmetic and manipulative operations. The JOBLIST system uses the SOLID systems to retrieve and update information in the libraries.

The SOLID system manages information and, in particular, processes both retrieval and update requests received from the JOBLIST system. It has been fully described ([1] and in the references quoted therein). Its mathematical basis is the JOBLIST data structure. It has three interdependent files (REG-FILE, AFILE and MFILE) that can be accessed independently of one another. MFILE contains the referenced information in compressed form. AFILE is a keyed-entry file that yields the so-called machine addresses to the compressed items of information in the MFILE. REGFILE is a simulated communications network whose "information paths" are described by the queries. This versatile file structure and performance data for the SOLID system have been described [1, 16]. A prototype of the SOLID system, coded in IBM Basic Assembler, was used as a part of the highly successful PENNRAMS system [17] for processing medical information.

The MATCHEM system is the actual alphanumeric deductive system. It can be viewed as a planned generalization of the CICLOPS [8] or EROS [6] versions of the mathematical model of Ugi and Dugundji [5, 7] for synthesis of organic compounds.

#### NUMERIC DEDUCTIVE SYSTEMS

The three interdependent systems in this category are CURFIT, CRAMS and FRANS. CURFIT is a subsystem for CRAMS, and CRAMS will be a subsystem for FRANS. However, the three systems may also be used on a stand-alone basis. Production versions of CURFIT and CRAMS, coded in FORTRAN IV, can be obtained from the author.\*

##### *The CURFIT system*

CURFIT is an automatic curve-fitting system that returns unambiguous, easily understood answers to questions about fits of data to user-specified

\*The cost (to cover mailing and reproduction) of the complete documentation of CURFIT and CRAMS is \$30 and \$40, respectively.

linear or non-linear equations [18, 19]. The methods used to test the reliability of curve-fitting methods have been described [20], and the use of CURFIT has been demonstrated [18–22]. Details of the implementation of CURFIT have been given in its Operation and Logic Manual [22]. Here it is sufficient to note that in the CURFIT system the reliability (or maximum tolerance) for every value is a part of the raw data. These maximum tolerances are used to reject spurious data points, to detect unsuspected curvature, and ultimately to determine the maximum errors that are associated with the computed median values for parameters. These maximum errors, which must not be confused with the probable errors of conventional curve-fitting methods, permit the direct comparison of either different equations fitted to the same data or the same equation fitted to different sets of data. There is no limit to the number of variables in the fitting equation. In that regard, CURFIT has been used to fit data to equations with 40 variables.

### *The CRAMS system*

The chemical reaction analysis and modeling system (CRAMS) is a general software system designed to handle reaction models with any combination of rate and equilibrium equations [23, 24]. Two different kinds of parameters can be identified. Variable parameters, like concentrations or observables, are normally multi-valued but they can also be single-valued (e.g., a constant concentration). Constant parameters, like rate constants, are normally single-valued but they can be multi-valued (e.g., for  $A \xrightarrow{k_1} B + C$  the values of  $k_1$  might be given by  $k_1 = k_0 \log(D/2(E - 0.1))$ , with  $D$  and  $E$  multi-valued parameters).

Two different kinds of questions can be asked about reaction models: prediction and computation questions. Prediction questions seek information about the various combinations of parameters for which data are required to compute values for any subset of the parameters and to test the validity of the proposed reaction model. Total solutions are those in which all the parameters are either given or computed. Partial solutions contain at least one parameter that is neither given nor computed. Partial solutions are not restricted to separately identifiable submodels of the reaction model. Obviously the predictive facility is of paramount importance because it can be used to help design experiments that will yield data for successful tests of the proposed reaction model. Computation questions result in the calculation of values for constant and/or variable parameters, and (if possible) tests of the validity of the model. Simulation questions are those in which only values for the variable parameters (e.g., concentrations) are computed. In processing computation questions, CRAMS first uses the predictive facility to determine the computable parameters and the order in which they must be computed. CRAMS has a pre-processing facility that permits the direct entry of raw data. The concentration data can be collected on different time scales. A post-processing facility permits users to manipulate and display both the original and the computed data in a variety of different ways.

The background literature has been fully reviewed [24, 26, 27]. The predictive facility of CRAMS [26, 28] appears to be unique. No other system appears to be capable of processing data for reaction models with any mixture of rate and equilibrium equations. Furthermore, there appears to be no other system that can process data collected on arbitrary time scales or offer the range of pre- and post-processing facilities that CRAMS does (see a comprehensive review [29]). Another apparently unique feature of CRAMS is the recently implemented, fully automatic error detection and corrective action (EDCA) system [25] that ensures the accuracy of computed results, and apparently eliminates the need to use special mathematical procedures to evaluate unstable functions. With respect to its computational facilities, CRAMS should be compared with the many specialized systems that can process only limited subsets of the computation questions that are routinely processed by CRAMS [24]. Examples are: (i) the many systems that can process only simulation questions about kinetic reaction models [30] and use special integration methods [31]; (ii) systems that can process questions for equilibrium models only [32–34]; and (iii) systems that use curve-fitting methods to compute reaction constants from data for the concentrations of reactants [35, 36]. There appears to be no single system or combination of systems that can perform all the tasks that are routinely done with CRAMS; examples are available [24].

Details of parts of CRAMS have been reported [23–27] particularly in the interface [23], user [24] and systems [25] manuals. The fully implemented final version will be reported in a later paper, as will the solution of some hitherto apparently intractable chemical/biochemical/chemical engineering problems (cf. [24]).

*Data structures and algorithms in the CRAMS system.* The fundamental concepts that have led to the successful realization of the CRAMS system include: (i) a user interface with extensive error recovery procedures that permits researchers to enter information in an easily understood familiar form; (ii) generalized, subject-independent data structures for internally representing input information and efficient algorithms for manipulating them; (iii) the concept that the kind of input information automatically determines both the computable parameters and the way in which they will be computed (the actual methods that are used to calculate computable parameters are not determined by the input); (iv) the idea that checking computed values is essentially a mechanical problem and can be fully automated. Moreover, the corrective actions that must be taken when a computational error is discovered are determined by the method of computation used and can, therefore, be automated also.

The computational procedures (iii) and (iv) have been discussed [25, 26]. These, together with the EDCA system, which is the implemented form of (iv), will be described in later papers, as will the generalized, subject-independent data structures and algorithms for manipulating them.

The five data structures used are constructed by the user interface when the chemistry-like description of the reaction model and its data are decoded.

First, the symbol table (SYTB) contains the names of all reactants, observables and reaction constants. Second, the FLUX matrix is a concise representation of all the individual chemical reactions in the model. Its elements are the stoichiometric ratios of reactants, with reactants on the left-hand side of reactions designated by minus signs. The rows and columns of the FLUX matrix correspond to reactants and reactions respectively. Rate reactions are always entered first. The status of every reactant and reaction constant is recorded in so-called status vectors: GIVEN, COMPUTE and RESULT.

The GIVEN vector describes the input status of every parameter. Three states are recognized: given ( $>0$ ); may be given or computed (0); and cannot be given ( $<0$ ). The COMPUTE vector is a record of the computational status of every parameter. It is constructed from the GIVEN vector and the FLUX matrix with the prediction algorithm [26]. Two states are recognized: computable ( $>0$ ) and not computable ( $<0$ ). The RESULT vector describes the output status of every parameter. It is constructed from the GIVEN and COMPUTE vectors. The six states that are recognized are: given (G), given and recomputed (R), computed (C), not given or computed (blank space), cannot be given and not computed (N), cannot be given and computed (NC).

*Modular structure.* CRAMS is a complex, highly modular, FORTRAN IV system that has about 120 routines in addition to approximately fifteen IBM system routines. There are plans to recode it in TPL [1] when that language is completed. The current version can be used on any IBM 360/370/3030

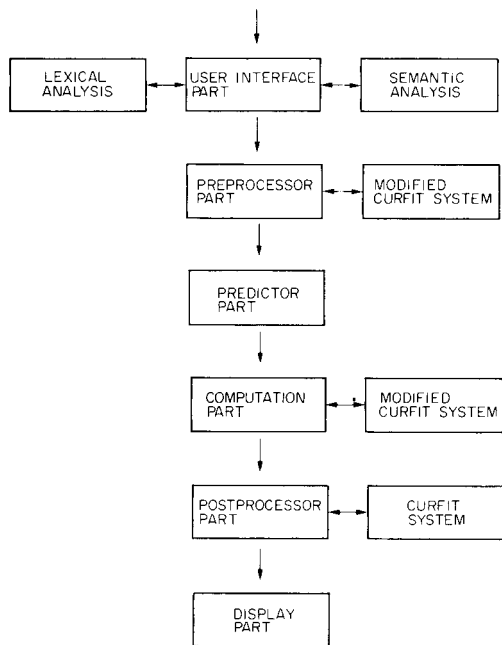


Fig. 2. Schematic flow chart for the CRAMS system.

machine that has 280K or more bytes of core memory. Here it is sufficient to note that the three local versions of CRAMS can handle models with up to 680 parameters.

A schematic flow chart for CRAMS is given in Fig. 2. The user interface accepts descriptions of the model and data in an easily understood chemistry-like language and converts them to the general data structures. In this part the functions that will be used later to process raw data are also constructed. The next, PREPROCESSOR, part is executed only for a computative problem (i.e., concentrations and/or constants are to be computed). It converts data collected on different time scales to a single compound scale and then computes the derivatives for every "given" concentration at every time.

In the PREDICTOR part, CRAMS operates in either the predictive or the computative mode to determine the combinations of data that are needed to solve for any subset of the parameters in the predictive mode or, in the computative mode, the parameters that can be computed; the minimum set of equations that must be solved; and the order in which they are to be solved.

The COMPUTATION part is executed only if parameters are to be calculated and it uses the information produced in the PREDICTOR part. It is in this part that the computed information is checked and, if there are computational errors, corrective actions are initiated. In the POSTPROCESSOR part the computed information and the input information (if necessary) are used to compute any new information that is requested. In the DISPLAY part, the original and computed information are displayed in an optional variety of ways as graphs, tabular form, etc. To a substantial degree the amount of information and the way that it is displayed is determined by the user.

### *The FRANS system*

The *function recognition and numerical solution*, FRANS, system is the logical extension of CRAMS that will process systems of one or more equations of virtually any type. There is to be no restriction on either the kind or combinations of equations. The FRANS system will have both a predictive and a computative capability and it will have an *error detection and corrective action* (EDCA) system that will assure that only reliable values for parameters that are computed are displayed.

The preliminary design for the FRANS system has been completed.

### EXAMPLES OF THE USE OF THE CRAMS SYSTEM

The CRAMS system can handle reaction models with any combination of equilibrium or rate equations. Within the available resources, there are no restrictions on either the size or complexity of the reaction model that can be processed. Many examples of its use have been described [24]. Here the simple model:

$A \rightarrow 2.1 B + C$ , RKF1;  $C + D \rightleftharpoons A$ , RKF2, RKB;  $X + B \rightleftharpoons C + A$ , EK;

is used to illustrate its use. The second equation is the reversible rate reaction:



### *Prediction problem*

To obtain all possible predictions for the above model, the input deck is as indicated in Table 2. The pertinent part of the output is displayed in Table 3. There are forty-seven different combinations of data that will yield complete solutions and there are three combinations that yield partial solutions (1, 2 and 22). It should be noted that not all possible combinations of data will yield solutions. For example, concentration data for A, C and X will not yield a solution.

### *Computation problem*

To illustrate the use of CRAMS, suppose that the concentrations for all reactants are to be computed at the times indicated below; the initial concentration of A is 0.1111; and the reaction constants are RKF1 = 10, RKF2 = 0.1, RKB = 5 and EK = 0.022 (Prediction 50 in Table 3.) The input deck is indicated in Table 4. The pertinent part of the self-explanatory output is displayed in Table 5.

The author is indebted to the many students and especially Dr. R. A. Butler, Dr. D. E. Whitten, Mr. J. A. Lucas II and Mr. M. Stubican for their contributions in designing and implementing the CRAMS system.

TABLE 2

Input deck for the prediction problem<sup>a</sup>

---

```
//EXEC CRAMSM
//INPUT. DATA DD *
/: SAMPLE PREDICTION PROBLEM:/
SYSTEM:
  SELECT = 1;
EQUATIONS:
  A → 2.1*B + C, RKF1; C + D ⇌ A, RKF2, RKB;
/: EQUILIBRIUM REACTIONS ARE ALWAYS LAST:/
  X + B ⇌ A + C, EK;
CONSTANTS:
/: ALL REACTANTS AND REACTION CONSTANTS MAY BE GIVEN:/
/: OR COMPUTED:/
  A = 0; B = 0; C = 0; D = 0; X = 0; RKF1 = 0; RKF2 = 0; RKB = 0; EK = 0;
STOP:
```

---

<sup>a</sup>Comments begin with /: and end with :/. SELECT = 1 means that it is a prediction problem. The remainder of the model description is self-explanatory.

TABLE 3

All possible predictions for the reaction model:  $A \rightarrow 2.1 * B + C$ , RKF1;  $C + D \rightleftharpoons A$  RKF2, RKB;  $X + B \rightleftharpoons C + A$ , EK. The prediction with the lowest multiplicity will in general yield the most accurate values for the computed reaction constants. Thus prediction 28 will generally yield the most accurate values for RKF1, RKF2, RKB and EK.

(The predictions that were output by the CRAMS system are to be tabled next. G—given; R—recomputed; BLANK—not computed; N—cannot be given; NC—cannot be given and computed. The multiplicities of given reactants are designated by N: N\*G. The prediction with the lowest multiplicity generally yields the lowest maximum errors for the reaction constants when they are computed with the CURFIT system. The predictor command (PRECOM) is 0 and the computational status of no variable was specified.)

## 30 NON-REDUNDANT PREDICTIONS ARE TABLED NEXT

NAME	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
A	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	3*G 3*G 3*G	
B	1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	
C	3*G 3*G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
D	C	2*G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
X	C	C 0*G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
RKF1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
RKF2	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
RKB	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
EK	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30		

## 20 NON-REDUNDANT PREDICTIONS ARE TABLED NEXT

NAME	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
A	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
B	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G	1*G 1*G
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
D	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
X	0*G 0*G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
RKF1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
RKF2	G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
RKB	G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
EK	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	

OUT OF 59 ORIGINAL PREDICTIONS 50 WERE NOT REDUNDANT

TABLE 4

Input deck for the computation problem<sup>a</sup>


---

```
//EXEC CRAMSM
//INPUT. DATA DD *
/:SAMPLE COMPUTATION WITH CRAMS:/
SYSTEM:
EQUATIONS:
  A → 2.1*B + C, RKF1;
  C + D ⇌ A RKF2, RKB; X + B ⇌ A + C, EK;
CONSTANTS:
  RKF1 = 10; RKF2 = 0.1; RKB = 5; EK = 0.022;
INITIAL CONCENTRATIONS:
  A = 0.1111;
DATA:TIME;
0.25 0.50 1.0 1.05 1.85 3 4 5 10 20 50;
STOP:
```

---

<sup>a</sup>Initial concentrations default to zero. So those for B, C, D and X equal zero.

TABLE 5

The pertinent part of the output for prediction 50 in Table 2 for the reaction system:  $A \rightarrow 2.1*B + C$ , RKF1;  $C + D \rightleftharpoons A$ , RKF2, RKB;  $X + B \rightleftharpoons C + A$ , EK. The initial concentrations are  $A = 0.1111$ ,  $B = C = D = X = 0$  and the reaction constants are  $RKF1 = 10$ ,  $RKF2 = 0.1$ ,  $RKB = 5$  and  $EK = 0.022$ .

(Tabulated results output by the SOLVER part of the CRAMS system. Four significant figures are to be printed for each of the values for the 5 compounds for those of the 12 data points that were computed. TRUNCATE, EQTOL, ACCURACY and MINSTEP are system parameters. For data point 1, TRUNCATE = 0, EQTOL =  $0.1000D - 05$ , ACCURACY =  $0.1000D - 01$ . For data points 2 to 12, TRUNCATE = 2, EQTOL =  $0.1000D - 05$ , ACCURACY =  $0.1000D - 0.1$ . The default values are: TRUNCATE = 0, EQTOL =  $0.1000D - 04$ , ACCURACY =  $0.1000D - 01$ , MINSTEP =  $0.1000D - 14$ . The minimum positive value for a concentration that was used for checking is  $0.1000D - 14$ . All negative concentrations greater than  $-0.1000D - 11$  were set to zero. All 12 of the requested data points were computed.)

---

TIME	A	B	C	D	X
0.2500D + 00	0.1111D + 00	0.0	0.0	0.0	0.0
0.5000D + 00	0.3671D - 02	0.1351D + 00	0.3096D - 01	0.2306D - 01	0.3824D - 01
0.1000D + 01	0.1319D - 02	0.1438D + 00	0.5986D - 01	0.2823D - 01	0.2496D - 01
0.1050D + 01	0.1240D - 02	0.1443D + 00	0.6169D - 01	0.2854D - 01	0.2409D - 01
0.1850D + 01	0.5943D - 03	0.1492D + 00	0.8113D - 01	0.3177D - 01	0.1469D - 01
0.3000D + 01	0.2897D - 03	0.1527D + 00	0.9451D - 01	0.3383D - 01	0.8150D - 02
0.4000D + 01	0.1759D - 03	0.1545D + 00	0.1005D + 00	0.3462D - 01	0.5201D - 02
0.5000D + 01	0.1148D - 03	0.1558D + 00	0.1040D + 00	0.3498D - 01	0.3483D - 02
0.7000D + 01	0.5881D - 04	0.1576D + 00	0.1074D + 00	0.3506D - 01	0.1822D - 02
0.1000D + 02	0.3341D - 04	0.1595D + 00	0.1090D + 00	0.3457D - 01	0.1038D - 02
0.2000D + 02	0.2412D - 04	0.1648D + 00	0.1096D + 00	0.3224D - 01	0.7294D - 03
0.5000D + 02	0.1937D - 04	0.1782D + 00	0.1100D + 00	0.2594D - 01	0.5435D - 03

---

THE PLOT COMMAND, PLOT = 0, INDICATES THAT NO PLOTS ARE TO BE GIVEN. FINISHED PROCESSING THE OUTPUT OF SOLVER.

---



## REFERENCES

- 1 P. A. D. deMaine and D. E. Whitten, *Advances in Information Systems Science*, 7 (1978) 89.
- 2 D. E. Whitten, Ph.D. Thesis, The Pennsylvania State University, 1976.
- 3 D. E. Whitten and P. A. D. deMaine, *The TPL Programming Language*, Report No. II, Global Management Systems, Computer Science Department, The Pennsylvania State University, 1976.
- 4 D. E. Whitten and P. A. D. deMaine, *IEEE Trans. Software Engineering*, 1 (1975) 111; *Operations and Logic Manual for Portable FORTRAN (PFORTRAN)*, Report No. I, Global Management Systems, Computer Science Department, The Pennsylvania State University, 1975.
- 5 J. Dugundji and I. Ugi, *Topics Curr. Chem.*, 39 (1973) 21.
- 6 I. Ugi, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum and W. Schubert, in E. V. Ludena et al., (Ed.) *Computers in Chemical Research and Education*, Plenum, New York, 1977.
- 7 I. Ugi, Proc. International Summer School Data Processing in Chemistry 80, Rzeszow, Poland, August 1980.
- 8 I. Ugi et al., in W. T. Wipke et al. (Eds.) *Computer Representation and Manipulation of Chemical Information*, Wiley-Interscience, New York, 1974.
- 9 P. A. D. deMaine, *The INTEGRAL Family of Reversible Compressors*, Report No. 2, Computer Science Department, The Pennsylvania State University, 1972.
- 10 P. A. D. deMaine and G. K. Springer, United States Patent 3,656,178, April 11, 1972.
- 11 P. A. D. deMaine, K. Kloss and B. A. Marron, *Nat. Bur. Stds. Tech. Note* 413 (1967).
- 12 P. A. D. deMaine and B. A. Marron, *Comm. ACM*, 10 (1967) 711.
- 13 P. A. D. deMaine, G. K. Springer and G. M. Campbell, *Proc. Am. Soc. Inf. Sci.*, 5 (1968) 109.
- 14 B. A. Minnihan and P. A. D. deMaine, Report No. 5, *Automatic Systems for the Physical Sciences*, Computer Science Department, The Pennsylvania State University, 1979.
- 15 P. A. D. deMaine, Proc. Fourth International Conference on Computers in Chemical Research and Education, Novosibirsk, 1979, p. 198.
- 16 P. A. D. deMaine and D. E. Whitten, *Management Datamatics*, 2 (1975) 31.
- 17 K. C. O'Kane, P. A. D. deMaine and R. J. Hildebrand, *Management Datamatics*, 4 (1975) 139.
- 18 P. A. D. deMaine and G. K. Springer, *Management Informatics*, 3 (1974) 233.
- 19 P. A. D. deMaine, G. K. Springer and R. A. Mikelskas, *Comput. Chem.*, 2 (1978) 7.
- 20 P. A. D. deMaine, *Comput. Chem.*, 2 (1978) 1.
- 21 P. A. D. deMaine, *Comput. Chem.*, 2 (1978) 53.
- 22 P. A. D. deMaine, Report No. 2, *Automatic Systems for the Physical Sciences*, Computer Science Department, The Pennsylvania State University, 1976.
- 23 M. Stubican and P. A. D. deMaine, Report No. 4, *Automatic Systems for the Physical Sciences*, Computer Science Department, The Pennsylvania State University, 1976.
- 24 P. A. D. deMaine, Report No. 5, *Automatic Systems for the Physical Sciences*, Computer Science Department, The Pennsylvania State University, 1980.
- 25 P. A. D. deMaine, Report No. 6, *Automatic Systems for the Physical Sciences*, Computer Science Department, The Pennsylvania State University, 1980.
- 26 R. S. Butler and P. A. D. deMaine, *Topics Curr. Chem.*, 58 (1975) 39.
- 27 P. A. D. deMaine, J. A. Lucas II and M. Stubican, in E. V. Ludena et al. (Eds.) *Computers in Chemical Research and Education*, Plenum, New York, 1977, pp. 25-48.
- 28 R. S. Butler, Ph.D. Thesis, The Pennsylvania State University, 1974.
- 29 D. Garfinkel, L. Garfinkel, M. Pring, S. B. Green and B. Chance, *Ann. Rev. Biochem.* 39 (1970) 473.
- 30 See, e.g., D. Garfinkel, M. C. Kohn and M. J. Achs, *Am. J. Physiol.*, 237 (3), (1979) R153.

- 31 See, e.g., M. Pring, *J. Theor. Biol.*, 17 (1967) 421.
- 32 L. G. Sillèn and B. Warngvist, *Acta Chem. Scand.*, 22 (1968) 3032.
- 33 E. C. Deland, Memo RM-5404-PR, Rand Corp. Santa Monica, California, 1967.
- 34 M. Bos and H. Q. L. Meershoek, *Anal. Chim. Acta*, 61 (1972) 185.
- 35 M. Pring, *J. Theor. Biol.*, 17 (1967) 430.
- 36 P. A. D. deMaine and R. D. Seawright, *Digital Computer Programs for Physical Chemistry*, Vol. II, MacMillan, New York, 1965.

## A PICTORIAL QUERY LANGUAGE FOR USE WITH ANY DATA BASE

CYNTHIA A. WALCZAK\*

*Microbial Systematics Section, National Institute of Dental Research, National Institutes Health, Bethesda, Maryland 20205 (U.S.A.)*

BARRY E. JACOBS

*Department of Computer Science, University of Maryland, College Park, Maryland 20742 (U.S.A.)*

(Received 23rd January 1981)

### SUMMARY

The query language discussed can be used with any type of data base and requires neither familiarity with formal logic nor knowledge of the data-base organization. The system is illustrated by creating a data base for the *Lactobacillus* hierarchy from a reference manual.

The problem of uniform methods of access is a major one to users of data bases. An institution may have several data bases, each with its own data format and data manipulation language. There is a need for some means of allowing users to operate across these data bases without having to learn different data manipulation languages and formats. The uniform access mechanism should be easy to learn and remember for a large class of users without training in mathematics or computer science.

A partial solution to the uniform access problem, Query-By-Example (QBE), has been implemented by Zloof [1]. However, it operates on only one kind of data base, the relational.

The new query language discussed in this paper is called Generalized Query-By-Example (GQBE) and can be used with any existing data base, relational, hierarchial or network. (A good introduction to these three types of data bases has been given by Date [2].) GQBE is a user-oriented, non-procedural data manipulation language. With GQBE it is not necessary for the user to be familiar with formal logic or the organization of the data base to be queried, as is true with so many other data manipulation languages. To execute a query, the user simply "fills in" pictures of the data-base tables on a screen with examples of answers. This main advantage of GQBE was achieved by redefining data-base formats so that a single common format describes any type or combination of types of data bases. This new data base format is termed Database Logic [3].

To illustrate the common format of GQBE, a data base created from the

*Lactobacillus* section of Bergey's Manual [4] is used. Bergey's Manual consists of about 1200 pages of verbal descriptions of microorganisms and incomplete keys. Typically, a search can be restricted to 200–300 pages. However, the organization of this reference manual does not lend itself to efficient searching. This is a problem with descriptive manuals all through microbiology. Information on the genus *Lactobacillus* in Bergey's Manual is organized into a hierarchy. Genus is at the top level and species and subspecies are at the bottom level, as can be seen in Fig. 1. (For simplicity, subgenus III is not included here.)

This hierarchy is next expressed as a GQBE data-base scheme. A GQBE data-base scheme is a collection of rules with table names to the left of the equals sign and the contents of the table, its column names, in parentheses on the right side of the equals sign. The *Lactobacillus* hierarchy becomes:

LACTOBACILLUS = (SUBGENUS, PRIME-FEATURES),  
 PRIME-FEATURES = (LETTER, SEC-FEATURES),  
 SEC-FEATURES = (SEQ#, SPECIES),  
 SPECIES = (SP#, SPNAME, SUBSPECIES),  
 SUBSPECIES = (SS#, SSNAME).

LACTOBACILLUS, PRIME-FEATURES, SEC-FEATURES, SPECIES, and SUBSPECIES are all names of tables. SUBGENUS, LETTER, SEQ#, SP#, SPNAME, SS#, and SSNAME are elementary column names. The meaning of SUBGENUS, LETTER, and SEQ# is explained in the legend to Fig. 1. SP# stands for species number, SPNAME stands for species name, SS# stands for subspecies number and SSNAME stands for subspecies name. Some names appear on the left side of one rule and on the right side of

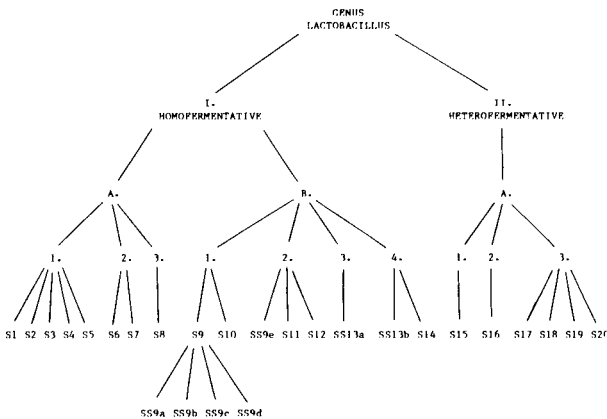


Fig. 1. The *Lactobacillus* hierarchy of Bergey's Manual. GENUS is at the top level, SPECIES, beginning with S, and SUBSPECIES, beginning with SS, are at the bottom level. There are no names for the intermediate levels in the Bergey's key. The following convention is used here: SUBGENUS corresponds to I and II, LETTER corresponds to A and B, and SEQ# corresponds to 1, 2, 3, and 4.

another rule. These are tables within tables. For example, LACTOBACILLUS is a table composed of two columns, SUBGENUS, and PRIME-FEATURES (Fig. 2a). But PRIME-FEATURES is itself a table composed of two columns, LETTER and SEC-FEATURES. There may be any number of columns in a table, for example, the SPECIES table has three columns (Fig. 2b). These tables can be merged by matching table names: The SUBSPECIES table is placed inside the SPECIES table. This SPECIES table is placed inside the SEC-FEATURES table, and so on until there is one table (Fig. 2c). The GQBE data-base table for subgenera I and II is shown in Fig. 3.

Two common search problems are to discover the properties of an organism, given its name, and to determine candidate organisms given information on several tests.

The properties or tests performed on microorganisms to classify them form a second data base in the Bergey's key, separate from that of the hierarchy of organism names. For illustrative purposes a small number of "key" features was chosen; those considered to be most reliable for differentiation within the genus *Lactobacillus* were used. These features and their abbreviations are: 15C, growth at 15 degrees centigrade; 45C, growth at 45 degrees centigrade; RIB, ribose fermented; CO<sub>2</sub>, carbon dioxide produced from glucose.

The same features are often found at many levels of the hierarchy. Figure 4 represents part of this network structure of the feature information. The center row lists the chosen features. The symbol ~ preceding a feature denotes absence of the feature. For example, ~RIB means Ribose is not fermented. Two levels of the *Lactobacillus* hierarchy are represented in Fig. 4. The top row represents the prime features level; the bottom row represents secondary features. Note that items at the secondary features level possess prime features also. For example, any items at the secondary features level beginning with "IB..." namely, IB1, IB2, and IB4, have all the features of IB at the prime features level as well as some features unique to themselves. Note also that another level is needed to distinguish between IB1 and IB2 (namely the lactic acid configuration produced at the species level). Figure 5 shows the GQBE description of the expanded network structure.

Now that the *Lactobacillus* data base has been defined for GQBE, the retrieval, update, insert, and delete operations can be illustrated by defining

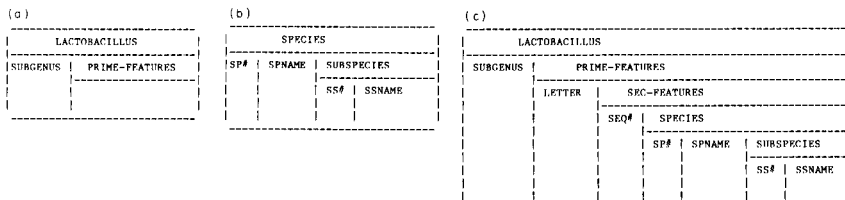


Fig. 2. Data-base structure of the table at (a) the genus level; (b) the species level; (c) the complete structure for genus *Lactobacillus*.

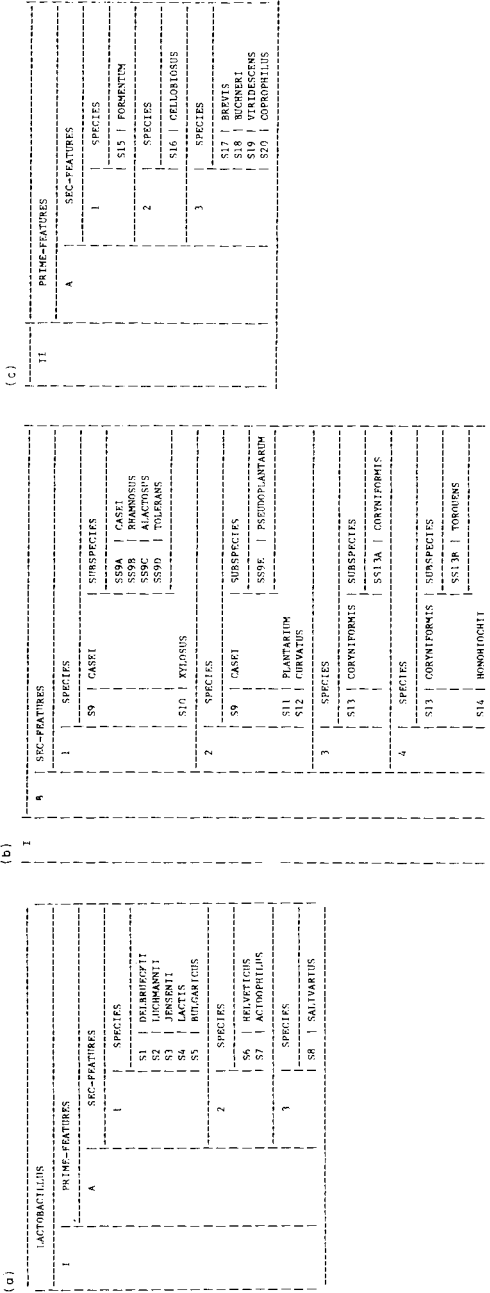


Fig. 3. The QBE data-base structure for the *Lactobacillus* hierarchy: (a) first part; (b) second part; (c) third part.

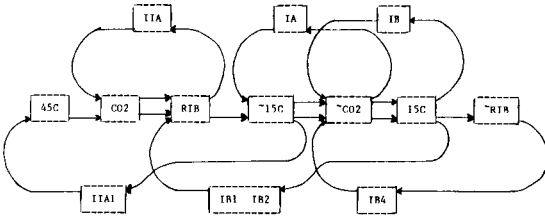


Fig. 4. Partial network structure of selected features for *Lactobacillus*.

CO2		RFB		45C		15C	
+	-	+	-	+	-	+	-
ITA1	IA1	IB1	IR4	ITA1	ITA3	IB1	IA1
ITA2	IA2	IB2				IB2	IA2
ITA3	IA3	ITA1				IB3	IA3
	IR1	ITA2				IR4	ITA1
	IR2	ITA3				ITA3	
	IB3						
	IR4						

LACTIC ACID		
D(-)	L(+)	DL
IA1	IA3	IA2
IB4	IR1	IR2
		IR3
		ITA1
		ITA2
		ITA3

Fig. 5. GQBE data-base structure for the expanded network of features partially shown in Fig. 4.

OP BOX		FEATURES			LACTOBACILLUS				
GET SP#, SPNAME, LACTIC ACID		RFB	15C	LACTIC ACID	SUBGENUS	PRIME-FEATURES			
		+	-	+	-	LETTER	SEQ#	SPECIES	
		ITA1		ITA1		A		SP#	SPNAME
							I	S9	CASEI

RESULT			
SP#	SPNAME	LACTIC ACID	
		D(-)	DL
S10	XYLOSUS	IR1	
S11	PLANTARUM		IR2
S12	CURVATIS		IR2
S17	BREVIS		ITA3
S18	HUCHWELI		ITA3
S19	VIRIDESCENS		ITA3
S20	COPROPHILUS		ITA3

Fig. 6. An example of retrieval.

the general format for a GQBE query. Every query is composed of two parts (see, e.g., Fig. 6). The OP BOX contains an operation keyword and the names of the tables and/or columns the user wants printed (the search results). The second part is a series of table skeletons from the data base in which the user specifies the search conditions.

*Retrieval*

The first example considers retrieval of species numbers and species names and types of lactic acid produced for species that grow at 15°C and ferment ribose. The GQBE query is shown in Fig. 6. The keyword for retrieval, GET, is placed in the OP BOX along with the table or column names the user wishes to see as the query result, in this case, SP#, SPNAME, and LACTIC ACID. Next, the tables that specify the conditions of the search are filled

in. In the FEATURES table, the user places an example element "IA1" under the positive column for features RIB and 15C and under the entire LACTIC ACID table. In the LACTOBACILLUS table, the SUBGENERA, LETTER, and SEQ# components of "IA1" are placed under their appropriate column headings. These serve to link the FEATURES table to the species number and name. S9 and CASEI are examples of species number and name.

The result of the query is a three-column table containing SP#, SPNAME of those organisms that grow at 15°C and ferment ribose and their corresponding entries in the LACTIC ACID table.

### Insertion

Not much is known about the species in *Lactobacillus* subgenus III. Suppose more tests were done on species S24 desidiosus and it was found to belong to IIA3. It is desired to insert species S24 in its appropriate place. In GQBE, this is accomplished as shown in Fig. 7.

Insertion is a two-step process. First the table that is to receive the new information must be retrieved; this is done with a GET instruction for the SPECIES table with search criteria SUBGENUS = II, LETTER = A, and SEQ# = 3. Here, II, A, and 3 are constants, not examples, so they are underscored. The result of this GET operation is that the SPECIES table for IIA3 is retrieved. The second step in inserting new information is to place INSERT in the second OP BOX and fill in the SPECIES table skeleton with new information.

### Update

Desidiosus was spelled incorrectly in Fig. 7. To correct information in the GQBE data base, the UPDATE operation is used. Like INSERT, the UPDATE operation is performed in two steps (Fig. 8). First the erroneous row of the appropriate table must be retrieved by means of the GET instruction. In the INSERT example above, the species table was retrieved for constants IIA3. Here only the newly inserted row of that species table is

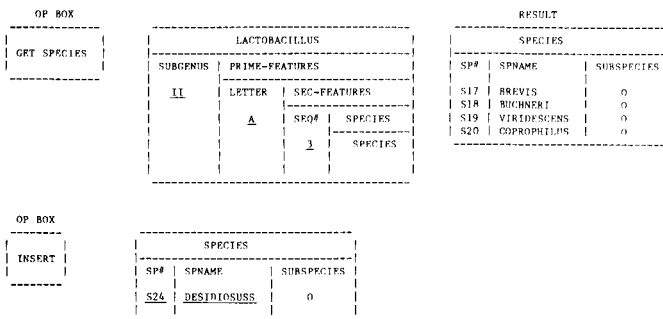


Fig. 7. An example of insertion.



OP BOX		LACTOBACILLUS					RESULT				
GET SPECIES		SUBGENUS	PRIME-FEATURES					SPECIES			
		<u>IL</u>	LETTER	SRC-FEATURES					SP#	SPNAME	SUBSPECIES
			<u>A</u>	SEQ#	SPECIES						
				<u>3</u>	SP#	SPNAME	SUBSPECIES				
					<u>S24</u>	DESID	SSP		S24	DESIDIOSUS	0

OP BOX		SPECIES		
UPDATE		SP#	SPNAME	SUBSPECIES
		S24	DESIDIOSUS	0

Fig. 8. An example of an update.

needed and so S24 is specified additionally as a constant in species table IIA3. Since the spelling of desidiosus is uncertain, the first five letters are used as an example element. SSP is an example of the subspecies table, also to be retrieved. The result of this GET operation is that row S24 of species table IIA3 is retrieved. In the second step of the UPDATE operation, "UPDATE" is placed in the second OP BOX, and the user changes the erroneous data elements.

### Deletion

The last example is of the delete operation: every occurrence of species S9 CASEI is to be deleted from the *Lactobacillus* data base (Fig. 9). The first step in the delete operation is to retrieve all rows of the species tables that contain S9 CASEI. So once again, GET is the operation keyword placed in the OP BOX with the column names of the desired results. In the skeleton of the LACTOBACILLUS table, "S9" and "CASEI" are underlined because they are constants. Sample elements for SUBGENUS, LETTER, and SEQ# were omitted, though they could have been included. Either way the search

OP BOX		LACTOBACILLUS					RESULT					
GET SPECIES		SUBGENUS	PRIME-FEATURES					SPECIES				
			LETTER	SRC-FEATURES					SP#	SPNAME	SUBSPECIES	
				SEQ#	SPECIES							
					SP#	SPNAME	SUBSPECIES					
					<u>S9</u>	<u>CASEI</u>	SSP		S9	CASEI	SS#	SSNAME
											SS9A	CASEI
											SS9B	RHAMNOSUS
											SS9C	ALACTOSUS
											SS9D	TOLERANS
									S9	CASEI	SS#	SSNAME
											SS9E	PSEUDOPLANTARUM

OP BOX		SPECIES		
DELETE		SP#	SPNAME	SUBSPECIES
		S9	CASEI	SSP

Fig. 9. An example of a deletion.

will range over the whole LACTOBACILLUS table. The second step is to specify "DELETE" in the second OP BOX. The species tables retrieved in the previous step are the search criteria. Notice that since the subspecies table is specified by the example element "SSP", all subspecies tables linked to "S9 CASEI" species tables are also deleted.

Of course, allowing unrestricted deletions can cause havoc in the data base. The power of GQBE is unrestricted. The data-base constructor must decide the limits of power given to the user. This is done by means of integrity constraints. Each time an insert, update, or delete request is received, it is first checked against the integrity constraints and is executed only if there are no violations. These considerations are discussed in detail elsewhere [5].

These examples demonstrate that Generalized Query-By-Example should be easy to use and would not require the user to know whether the underlying data-base is hierarchic, network, or relational. Further, it will free the constructor of data bases from concern about the requirements of the query language. Each part of the data base can be built in a natural manner as judged by the user, not the computer.

#### REFERENCES

- 1 M. M. Zloof, Query-by-Example: a database language, IBM Systems Journal, 16 (1977) 324.
- 2 C. J. Date, An Introduction to Database Systems, Addison-Wesley, Reading, MA, 2nd edn., 1977.
- 3 B. E. Jacobs, On Database Logic, Department of Computer Science Technical Report TR 737, Univ. of Maryland, 1978.
- 4 M. Rogosa, in R. E. Buchanan and N. E. Gibbons (Eds.), *Lactobacillus* group of bacteria, Bergey's Manual of Determinate Bacteriology, 8th edn., Williams & Wilkins, Baltimore, MD, 1974, pp. 577-579.
- 5 B. E. Jacobs and C. A. Walczak. A Generalized Query-by-Example Data Manipulation Language Based on Database Logic, Dept. of Computer Science Tech. Rep. TR914, Univ. of Maryland, 1980.

## A DATA BASE OF DATA BANKS FOR TOXICOLOGICAL INFORMATION

TSUGUCHIKA KAMINUMA

*The Tokyo Metropolitan Institute of Medical Science 3-18-22 Honkomagome, Bunkyo-ku, Tokyo 113 (Japan)*

AKIHIRO KURIHARA

*Research Organization for Disease Control 3-18-22 Honkomagome, Bunkyo-ku, Tokyo 113 (Japan)*

(Received 23rd January 1981)

### SUMMARY

The procedures necessary to find the appropriate data banks in seeking particular information or data are much less systematic than the way in which the information is stored at some data banks. Based on the information taken from several hundred direct-mailed questionnaires, a conceptual design is proposed for a data base of toxicological data banks relating to other areas such as medicine, pharmacology, biology, chemistry and environmental science. The system (not yet implemented) contains nearly 150 data banks (both computerized and manual) all over the world with data on the type of information, the way to obtain it, its cost, etc.

Information on chemicals is of vital importance in modern society for preserving the quality of life and preventing chemical hazards. Many data banks and information systems have been proposed for gathering, storing and disseminating chemical information and, since toxic substances control laws have been established in many countries, some attempts have been made to develop toxicology information systems in these countries.

Toxicology is a highly inter-disciplinary area. The data resources can be classified into several categories: computerized in-house systems, intra-national service systems, inter-national service systems, and un-computerized manuals and documents. Depending on the mode of access, these data resources may also be classified into direct and indirect service systems. In direct service systems, the user can himself gain access to the data, while in the indirect service system the service personnel must seek the information needed. With regard to existing information systems, the general user should know how many data bases are available on a specific topic and how a list of information resources, in any form, on the subject can be obtained. Further, an assessment of the real usefulness of these data banks is needed, as well as information on how the service systems can be used and their costs. If there is only one service channel for each source of information then matters

may be simple, but the situation is becoming complicated by the fact that multiple service channels for obtaining a certain piece of information are available. Unfortunately, there is no information system which gives such comparative information on toxicology for general users.

Thus the purpose of the present study was two-fold: first, to develop an information system, or more specifically, a data base for data banks in the area of toxicology; second, to use it to ally such data banks so that they are organized into an integrated information network. The work is not yet complete, and only preliminary results are presented below.

## TOXICOLOGICAL INFORMATION SYSTEM

The proposed system for toxicological information contains a data base of data banks as its central component. The purposes of the system are: (1) to predict and to give warning of a wide range of chemical hazards; (2) to monitor and survey chemicals in the biosphere; (3) to help people to take appropriate action after accidents; and (4) to help research for controlling chemicals, and to provide appropriate information for public dissemination.

The system should be useful to administrative agencies, chemical, drug and other industries, research and educational institutions, consumer groups, etc.

Figure 1 shows the block diagram of the system tentatively called TOXIN (Toxicological information network) [1]. Three independent systems are loosely coupled either directly or indirectly in TOXIN: the core data bases, the peripheral data banks, and the research support system.

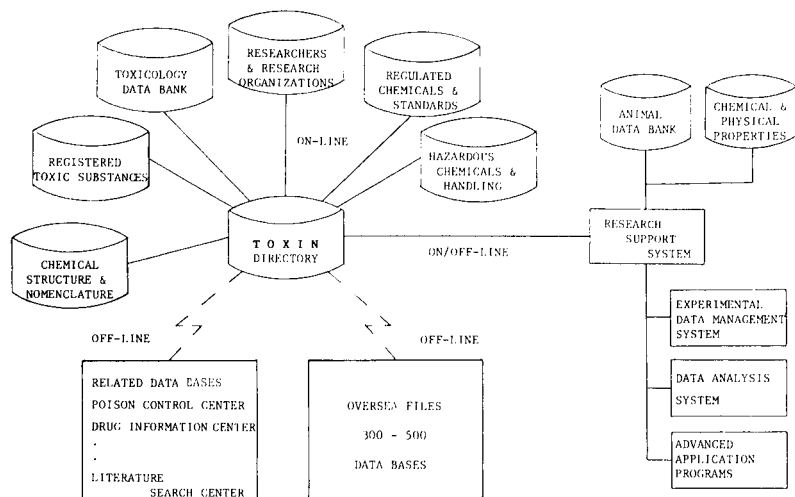


Fig. 1. The TOXIN system.

The core data bases are the data bases which are directly linked by communication lines. The core data bases consist of six categories of information files: (1) the file used for identifying chemical substances; (2) and (3) files containing information on biological effects (acute and chronic toxicity) of chemicals; (4) files of researchers and research organizations; (5) files of regulations and standards; (6) files containing information on the hazards and handling of chemicals.

The peripheral data banks include both computerized and manual information files relevant to the goal of the system and should cover international as well as intra-national toxicological data resources. The research support system contains data bases, an experimental data management system, a data analysis system, and advanced application programs for helping researchers at their laboratory.

Among the data banks in TOXIN, only those which are included in the core data bases can be retrieved on-line. Peripheral data banks must be searched independently after the user has obtained information on their usage through the TOXIN directory. Thus the TOXIN directory which couples the three systems acts as a switchboard for the core data bases and as an information directory for the peripheral data banks.

### *Realization of TOXIN*

As Fig. 1 indicates, TOXIN is a very pragmatic system. In fact, many of the data bases and application programs in TOXIN already exist elsewhere. The main purpose of TOXIN is to develop a mechanism, the data bank directory system, for associating existing components so that they become serviceable on request. The only TOXIN components which do not exist at present are the (Japanese) regulated chemical standard, the (Japanese) researcher organization file, and the TOXIN directory.

The present Japanese laws regulating chemicals standards make it necessary to include seventeen laws in the file. Updating of this information must be done by scanning the weekly government official reports (Kanpo in Japanese). Table 1 lists some source lists of chemicals published by Japanese government departments. The difficulty of computerizing these data is the specification of certain groups of chemicals by ambiguous group names, e.g. alkyls and acyls. Except for this problem, computerization of the regulated (Japanese) chemical standards is quite straightforward.

The data sheet for the researcher and researcher organization file has been designed. The purpose of this file is to associate chemicals with researchers and research organizations. There may be only a limited number of specialists on certain chemicals. It is vitally important to find these specialists or their affiliated institutions for urgent toxicological actions.

Table 2 lists the candidate data bases for other TOXIN core data bases. Even though TOXIN does not have any revolutionary concept, its realization will not be easy. Two major obstacles are the high cost and bureaucratic sectionalism. There is no easy solution for these problems. However, the

TABLE 1

## Publication of regulated chemical substances

Lists of chemicals	Regulations	Agency
List of existing chemical substances	Chemical Control Law	Ministry of International Trade and Industry
List of existing chemical substances	Industrial Safety and Health Law	Ministry of Labour
Priority list for assessment of existing chemicals in the environment		Environment Agency
Handbook for chemical reaction hazards		Tokyo Fire Department

success of CIS project [2] shows that even a small system can grow into a large influential international system if its development and operation philosophy are adequate. Therefore, the practical strategy for realization of TOXIN is first to develop a pilot system which can be started without external help. This pilot system is designed for research workers, who will probably be the first and main users of TOXIN.

*Pilot system for TOXIN*

Figure 2 shows the pilot system for TOXIN. It consists of two subsystems: the CISC (chemical information system complex) and an SWS (scientific work station for chemists). Both the CISC and the SWS are designed to support research works in toxicology or in chemical information. The CISC includes data bases and application programs which are best provided at a common computing facility, while the SWS consists of smaller data bases, data management systems, and application programs suitable for provision at a local computing facility. The CISC will be linked to more than one SWS, and a SWS operator can access other data banks as well (Fig. 3). Some SWS are linked to a laboratory data acquisition system.

In this pilot system, the central control unit plays the same role as the TOXIN directory. Figures 4 and 5 show block diagrams of the central control unit and its function, respectively.

## TOXICOLOGICAL DATA BANK DIRECTORY

In this section, the process and results of gathering information on toxicological data banks are discussed. Figure 6 shows the data sheet used for collecting information. The sheet was designed for manual recording of information, but was also used as the questionnaire for this survey. Initially, a format similar to that used in the Mitre Corporation report on a chemical substance data base [3] was considered but that sheet was later found to be inappropriate for the present purpose. Corresponding data sheets in Japanese are used for Japanese information sources.

TABLE 2

List of the candidate data bases for TOXIN core data bases

Subsystem	Data bases	Agency	File description
CHEMICAL STRUCTURE NOMENCLATURE	SANSS/CIS CHEMICAL NAME FILE	NIH-EPA/US NCI/DHW/US	SUBSTANCE'S IDENTIFICATION Substance Prime Name CAS Name, IUPAC Name, Synonyms, CAS Registry Number CHEMICAL STRUCTURE Connection Table Wiswesser Line Notation CHEMICAL & PHYSICAL PROPERTIES COMPOSITION
REGISTERED TOXIC SUBSTANCES	RTECS (Registry of Toxic Effects of Chemical Substances)	NIOSH/DHW/US	SUBSTANCE'S ID MOLECULAR WEIGHT MOLECULAR FORMULA TOXIC DOSE DATA Carcinogen, Neoplastigen, ... Species Exposed TLDo, TCLo, LDLo, LD <sub>50</sub> , LCLo, LC <sub>50</sub> , ...
TOXICOLOGY DATA BANK	TDB (Toxicology Data Bank) BDT (Data base in Toxicology)	NLM/DHW/US INSERM/FRANCE	REFERENCES SUBSTANCE'S ID TOXICOLOGICAL DATA Human Toxicity Excerpts Animal Toxicity Excerpts Interaction Excerpts Laboratory Methods Excerpts
RESEARCHERS & RESEARCH ORGANIZATIONS	TIMS-RESEARCHER FILE SSIE CLEARING FILE (Smithsonian Science Information Exchange) CLEARING FILE	TIMS/JAPAN SSIE/US JICST/JAPAN	RESEARCH ORGANIZATION RESEARCHER RESEARCH CLEARING Project Title, Project Summary, ...

TABLE 2 (continued)

Subsystem	Data bases	Agency	File description
REGULATED CHEMICALS & STANDARDS	LIST OF EXISTING CHEMICAL SUBSTANCES	MITI/JAPAN ML/JAPAN	SUBSTANCE'S ID REGULATIONS STANDARDS
HAZARDOUSNESS OF CHEMICALS	PRIORITY LIST FOR ASSESSMENT TSCA Inventory List OHM-TADS (Oil & Hazardous Materials — Technical Assistance Data System) HANDBOOK FOR CHEMICAL REACTION HAZARDS	EA/JAPAN EPA/US EPA/US  TOKYO FIRE DEPT./JAPAN	SUBSTANCE'S ID CHEMICAL PROPERTIES HANDLINGS ENVIRONMENTAL EFFECTS BIOLOGICAL EFFECTS



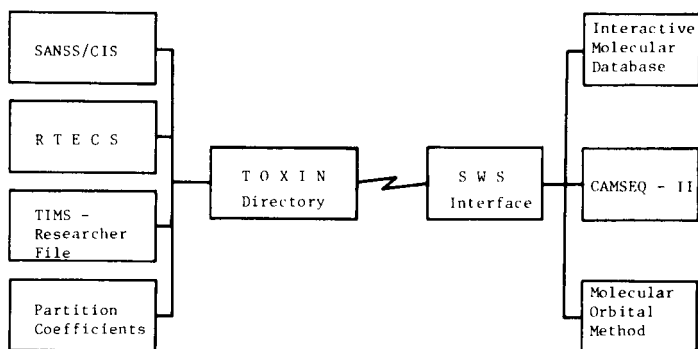


Fig. 2. The TOXIN pilot system.

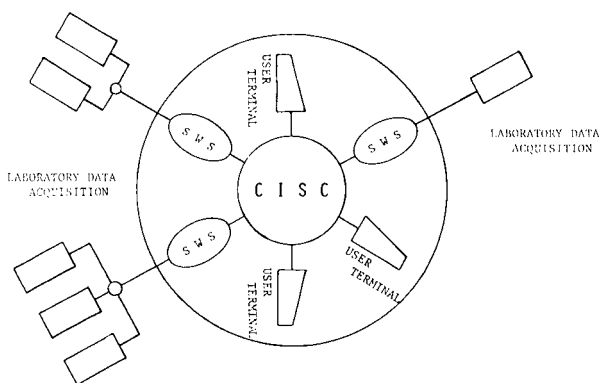


Fig. 3. CISC (chemical information system complex) concept.

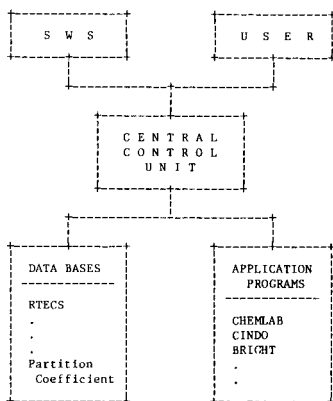


Fig. 4. Block diagram of the central control unit.

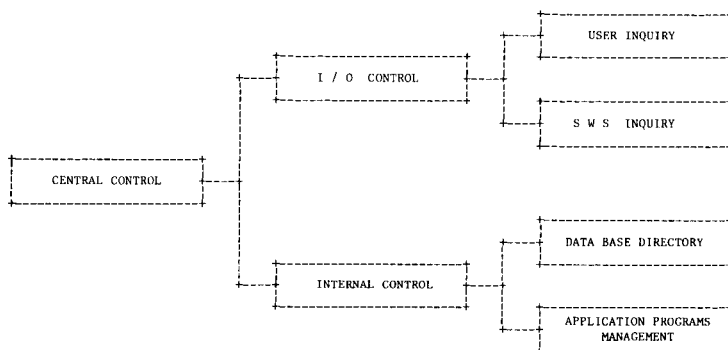


Fig. 5. Functions of central control unit.

The sheet is designed to provide practical information for the user such as where the information is, how to obtain it, how many different service channels and media there are, how much it will cost, etc. The back page of the sheet (Fig. 6B) contains information on the dissemination of the data banks: agencies and their addresses, on-line service systems, and information media.

Over 450 such questionnaires were mailed to 280 places outside Japan, and answers were received from 200 places, including government institutions, universities and industries. Table 3 lists the distribution of these correspondents over the nations. More than 80% in the list are in the United States. The total number of the data bases was 150, of which 28 are manual and 122 are computerized.

Among 122 computerized data bases, 31 are used in-house. Many information systems in the pharmaceutical industry are in this category. There are only four data bases whose use is restricted to a particular nation whereas 87 data bases are available for international use. Once a data base is used openly within a nation, it may also be used internationally.

The validity of the number of data bases that could have been covered may be questioned. Whether or not the survey could cover almost all existing toxicological information is difficult to establish. However, the Cudra association kindly investigated the present list of data bases, and found that only 4 data bases were unlisted in their directory of on-line data bases [4]; also the Mitre Corporation's report for the Chemical Substances Information Network [3] listed about 300 data banks, of which 80% were computerized. The present survey showed that many data banks listed in that report have already been discarded and that many are not operational. The latter fact indicates that updating and careful checks on the availability of data bases are unavoidable.

TABLE 3

List of distribution of correspondence

Nation	Collaborated organizations		Nation	Collaborated organizations	
	Dispatch	Receipt		Dispatch	Receipt
U.N.	2	2	Indonesia	1	1
W.H.O.	4	4	India	1	1
Austria	1	1	Hungary	3	4
Belgium	1	1	Luxemburg	—	1
Brazil	1	1	Netherlands	5	3
Bulgaria	2	2	South Africa	8	9
Canada	11	6	Spain	1	—
Czechoslovakia	3	1	Sweden	1	2
Denmark	1	1	Switzerland	3	3
Egypt	1	5	U.K.	16	15
France	13	8	U.S.A.	185	128
Germany	3	3	U.S.S.R.	5	1
			Total	280	200

## DISCUSSION

The first apparent application is to use the above information as a manual data bank of toxicological data banks. Annual surveys via the questionnaire will provide the basic information. Computerization of this information is straightforward, given suitable search software with appropriate key words. The second and main application will be to use the data as the TOXIN (or its pilot system CISC) directory. When the list of chemicals included in the data bases by their identity number (e.g., CAS registry number) is added, then an inverted file listing each chemical against the names of relevant data bases can be created. The original file of data banks and the inverted file provide enough information for the TOXIN and its pilot system. In fact, the SANSS of NIH/EPA CIS [2] is an example of this kind of directory system. The present plan is to divide the SANSS into two parts, namely the chemical substance dictionary and the inverted data base directory. The former is regarded as a peripheral component; only the directory of data bases with the inverted file are left at the central unit.

Development is currently proceeding along these lines with the DEC-20 system at the Fujimic computer center for the CISC, and the PDP 11/70 at TMS for the SWS machine.

## CHEMICAL INFORMATION RESOURCE

TIMS NUMBER \_\_\_\_\_

IDENTIFICATION: \_\_\_\_\_ AS OF \_\_\_\_\_

ACRONYM: \_\_\_\_\_

AGENCY: \_\_\_\_\_ DEPT. & DIVISION: \_\_\_\_\_

CONTACT: \_\_\_\_\_ CITY: \_\_\_\_\_

STREET ADDRESS: \_\_\_\_\_ STATE: \_\_\_\_\_ ZIP: \_\_\_\_\_ COUNTRY: \_\_\_\_\_

PHONE: \_\_\_\_\_

TYPE OF DATA SOURCE:  COMPUTERIZED  MANUAL  
 IN-HOUSE  NATIONAL USE ONLY  AVAILABLE ABROAD  
 (INTERNATIONALLY)

SIZE OF TOTAL DATA BASE: \_\_\_\_\_ (No. OF RECORDS)  
 \_\_\_\_\_ (No. OF BYTES)

ANTICIPATED GROWTH: \_\_\_\_\_ (No. OF CHEMICALS/YR)

NUMBER OF UNIQUE CHEMICALS: \_\_\_\_\_

CHRONOLOGICAL COVERAGE: \_\_\_\_\_ TO \_\_\_\_\_

FILE UPDATE FREQUENCY: \_\_\_\_\_

TYPE OF DATA IN SYSTEM:  RAW  
 ANALYZED  
 SUMMARIZED  
 BIBLIOGRAPHY  
 REFERRAL  
 OTHERS

-----

SUBSTANCE IDENTIFICATION:	BIOLOGICAL EFFECTS:	OTHER CONTENTS:
<input type="checkbox"/> MOLECULAR FORMULA	<input type="checkbox"/> BIOCHEMICAL STUDIES	<input type="checkbox"/> PRODUCTION ASPECTS
<input type="checkbox"/> CA REGISTRY NUMBER	<input type="checkbox"/> CLINICAL STUDIES	<input type="checkbox"/> MARKETING
<input type="checkbox"/> CAS NAME	<input type="checkbox"/> TOXICOLOGICAL STUDIES	<input type="checkbox"/> EXPOSURE
<input type="checkbox"/> IUPAC	<input type="checkbox"/> RODENTS	<input type="checkbox"/> EPIDEMIOLOGY DATA
<input type="checkbox"/> SYNONYMS	<input type="checkbox"/> ACUTE	<input type="checkbox"/> ENVIRONMENTAL EFFECTS
<input type="checkbox"/> WISWESSER LINE NOTATION	<input type="checkbox"/> WILDLIFE	<input type="checkbox"/> STANDARDS & REGULATIONS
<input type="checkbox"/> CHEMICAL STRUCTURE	<input type="checkbox"/> PRIMATES	<input type="checkbox"/> MANAGERIAL/
<input type="checkbox"/> CHEMICAL PROPERTIES	<input type="checkbox"/> IN VITRO	<input type="checkbox"/> ADMINISTRATIVE
<input type="checkbox"/> PHYSICAL PROPERTIES	<input type="checkbox"/> OTHERS	<input type="checkbox"/> SPECTROSCOPIC DATA
<input type="checkbox"/> COMPOSITION	<input type="checkbox"/> TERATO.	
<input type="checkbox"/> OTHERS	<input type="checkbox"/> GENICITY	
	<input type="checkbox"/> OTHER	
	<input type="checkbox"/> CHRONIC	

SYSTEM CHARACTERIZATION (200 Words or less - underline key words)

-----

[A]

## DISTRIBUTION:

1 PUBLICATION:

TITLE: \_\_\_\_\_ DATE OF ISSUE: \_\_\_\_\_

PUBLISHER: \_\_\_\_\_

OVERSEAS AGENT: \_\_\_\_\_

AGENT IN JAPAN: \_\_\_\_\_

PRICE: \$ \_\_\_\_\_ ¥ \_\_\_\_\_

2 MICROFICHE:

PUBLISHER: \_\_\_\_\_ DATE OF ISSUE: \_\_\_\_\_

OVERSEAS AGENT: \_\_\_\_\_

AGENT IN JAPAN: \_\_\_\_\_

PRICE: \$ \_\_\_\_\_ ¥ \_\_\_\_\_

3 MAGNETIC TAPE:

ORIGINAL DISTRIBUTOR: \_\_\_\_\_

SECONDARY DISTRIBUTOR: \_\_\_\_\_

PRICE: \$ \_\_\_\_\_ ¥ \_\_\_\_\_

4 ON-LINE SERVICE:

SYSTEM NAME: \_\_\_\_\_

5 OTHER: \_\_\_\_\_

MEDIUM: \_\_\_\_\_

CONTENT: \_\_\_\_\_

-----

OUTLINE OF CONTRACT: \_\_\_\_\_

-----

REFERENCES/REMARKS: \_\_\_\_\_

-----

TIMS USE ONLY: \_\_\_\_\_

[B]

Fig. 6. Data sheet for chemical information resource.

## REFERENCES

- 1 T. Kaminuma, S. Kurashina, T. Yamamoto and A. Kurihara, *The Chemical Information Symposium*, 6-2 (1979) 105 (in Japanese).
- 2 G. W. A. Milne and S. R. Heller, *ACS Symposium Series No. 54*, 1977, pp. 26-45.
- 3 M. Bracken, J. Dorigam, J. Hushow and J. Overbey II, *Chemical Substances Information Network*, The Mitre Corporation, McLean, VA, 1977.
- 4 *Directory of Online Data Bases*, Cuadra Associates, Inc., Santa Monica, CA, 1978.

## GRAPHICAL REPRESENTATION OF THE SOLUTIONS OF THE SCHRÖDINGER EQUATIONS FOR A PARTICLE IN VARIOUS MODEL POTENTIALS

HARUO HOSOYA\* and MAYUMI HIROSE

*Department of Chemistry, Ochanomizu University, Bunkyo-ku, Tokyo 112 (Japan)*

(Received 18th October 1980)

### SUMMARY

Several one-dimensional Schrödinger equations for a particle trapped in various model potentials are solved analytically or approximately, and the results are plotted by an X—Y plotter. Typical results are illustrated. Several interesting features of the relations between the shape of the potential and the solution of the Schrödinger equation are pointed out. Combination of these results allows reasonable guesses about the forms of the energy levels and wavefunctions for complicated but realistic problems.

Very few of the atomic and molecular Schrödinger equations can be solved rigorously, but exact solutions can sometimes be found for a particle trapped in a one-dimensional model potential. In some cases, students will obtain no idea of the close relationship between the shape of the potential and the solutions just by looking at the formulae of the exact solutions. This is also true for the application of the perturbation technique to simple systems.

In order to supply a good collection of materials for understanding the ideas of quantum mechanical reasoning, we have calculated and printed out through an X—Y-plotter a large number of the numerical solutions (rigorous or approximate) of the Schrödinger equations for a particle trapped in one-dimensional model potentials, namely, a square well (of infinite and finite depth), a knife edge potential, a V-shaped well, a champagne bottle-shaped well, a bell-shaped well (Eckart potential), etc. In cases where one or a few parameters are involved, their values were changed over a wide range so that several different features of the energy levels and wavefunctions become apparent.

Some of the potentials were found to be good models for realistic problems, such as the potentials of a heteropolar diatomic molecule, or a hydrogen-bonded bridge with double minima. In the latter case, the extent of the tunnelling effect can be well demonstrated in terms of the relative values of the parameters.

### METHOD OF CALCULATION

The solutions for the bound states of the one-dimensional Schrödinger equations

$$\left\{ -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) - E \right\} \psi(x) = 0 \quad (1)$$

with various shapes of the potential  $V(x)$  were enumerated by standard methods and plotted by the X-Y plotter equipped with the HITAC M-150 computer at Ochanomizu University. In the following discussions all the quantities become dimensionless by putting  $\hbar^2/2m = 1$ .

Some conventional subroutine programs such as the Bessel and hypergeometric functions involving series expansions were found to give erroneous values for extremely small and/or large arguments. Although details will not be given here, special precautions were taken to achieve "neat results".

## RESULTS AND DISCUSSIONS

### *Square-well potential with a finite depth*

The wavefunction  $\psi(x)$  takes either a sine or cosine curve within a finite square well with depth  $V_0$  and width  $l$ , and decays exponentially towards the outside [1-3]. A typical example is shown in Fig. 1. The number of bound states depends on the  $(V_0)^{1/2}l$  values as shown in Fig. 2. In Fig. 3 is shown the effect of the depth of the square well on the wavefunction with a given number of nodes. Although these results are well known, it is worthwhile for both students and research workers to study these figures to understand the nature of the wave equation.

### *Rectangular double minimum potential*

When the central part of the bottom of the square well is raised vertically, a rectangular double minimum potential is obtained. The solutions of the

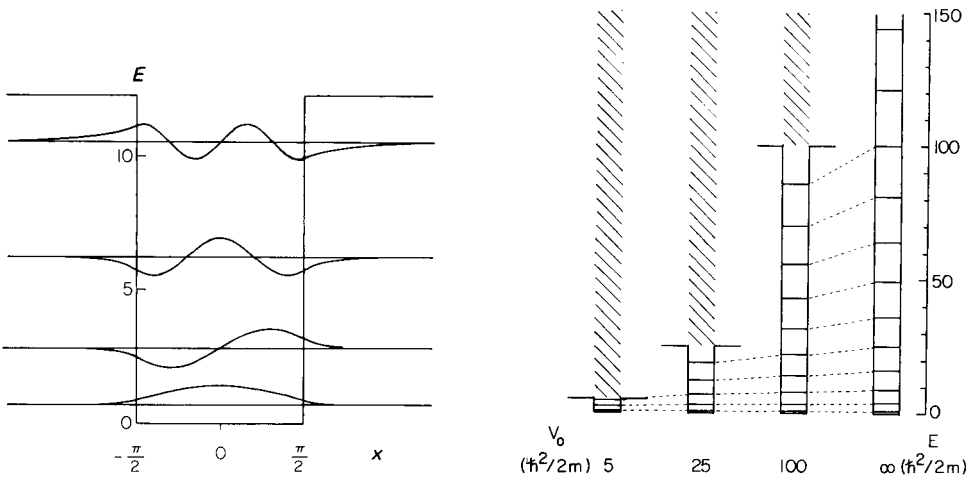


Fig. 1. Energy levels and wavefunctions of the square-well potential of finite depth,  $V(x) = V_0 = \hbar^2\gamma^2/2m$  with  $\gamma = 3.5$ . The width of the well is  $2l = \pi$ .

Fig. 2. Comparison of the energy levels for square-well potentials with various depth. The width of the well is kept constant as  $2l = \pi$ . The values of  $\gamma^2$  for  $V_0 = \hbar^2\gamma^2/2m$  are given. The number of the bound states is about  $[2l\gamma/\pi]$ .

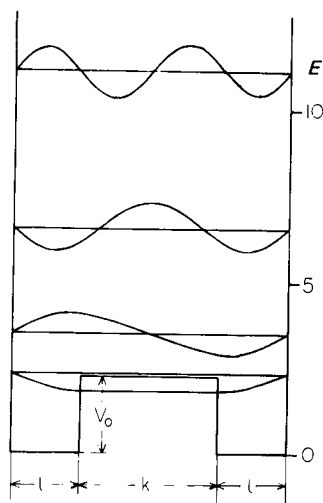
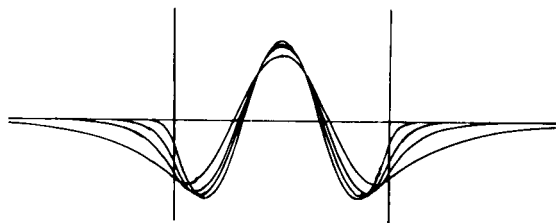


Fig. 3. The effect of the depth of the square well on the shape of the wavefunction with a given number of nodes. The values of  $\gamma$  corresponding to the curves from outside to inside are 2.5, 3.5, 5, and 10, respectively, and the width of the well is  $2l = \pi$  throughout.

Fig. 4. A rectangular double minimum potential with a low barrier.  $V_0 = \hbar^2 \gamma^2 / 2m$  with  $\gamma = 1.5$ , and  $k = 2$  and  $l = 1$ .

Schrödinger equation are obtained in a straightforward manner as in the former case [2, 3]. Two typical results are given in Figs. 4 and 5. In the former case, the lowest energy is just above the central potential wall. The shape of the fairly flat wavefunctions with large amplitude is to be noted. Also noteworthy in Fig. 4 is that the energy gap counting from the lowest energy level increases in the ratio of 1:3:5:..., as if the two "legs" of the rectangular double minimum potential were cut out to give a simple square well.

The results given in Fig. 5 illustrate other interesting features of the double minimum potential. The central wall is so high and thick that the lowest bound states are almost degenerate. However, the positions of the two maxima of the corresponding wavefunctions are pulled a little toward the center showing the effect of tunnelling. All the other wavefunctions, with energies higher than the central wall, have nodes near both sides of the infinite potential walls. By joining them with dashed lines, another square well (but with slightly slanted walls) appears as in the top half of Fig. 5. If one looks at the inside of this phantom potential together with the wavefunctions and energy levels already drawn for the original potential, the origin of the large amplitude of the wavefunction of the third lowest level above the thick wall in the central part of the well can be immediately understood. Further, by changing the height and width of the central wall, one can see how the wavefunction trapped in one of the legs tunnels through the central wall and joins with its counterpart in another leg. This type of potential can be used as a model potential for a symmetrical hydrogen bond bridge.



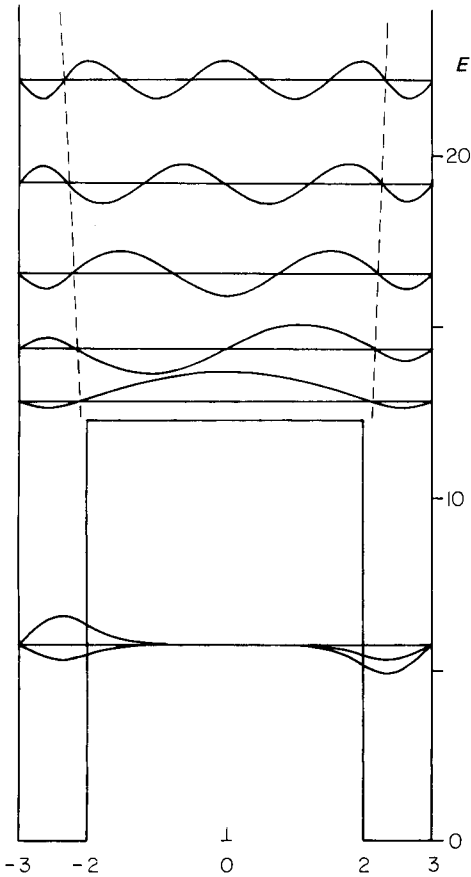


Fig. 5. A rectangular double minimum potential with a high thick barrier.  $\gamma = 3.5$ ,  $k = 4$ , and  $l = 1$ . Dashed lines are drawn to show a phantom square well above the central potential barrier. The two lowest states are almost degenerate.

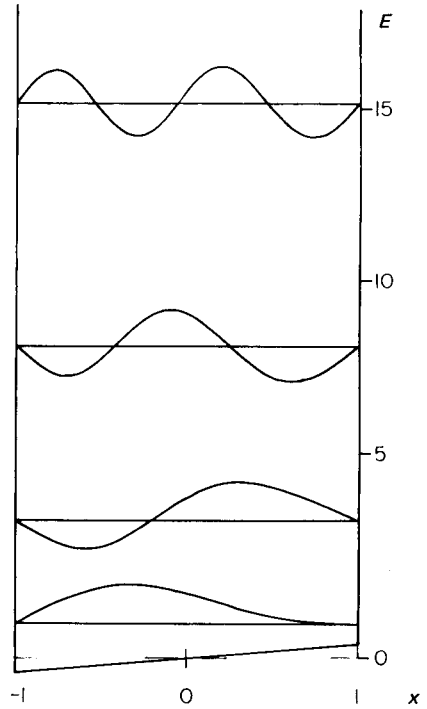


Fig. 6. A knife-edge potential.  $V(x) = bx$  (in units of  $\hbar^2/2m$ ) with  $b = 0.4$ .

It was also found that if the energy level is close to the roof of the central wall, a somewhat distorted wavefunction is obtained at the rectangular corner. This means that, as the wavefunction cannot follow the sudden change in potential, its second derivative does not become continuous.

### *Knife-edge potential*

If the slope of the knife-edge potential (Fig. 6) is not very large, the solution of the Schrödinger equation can be well approximated by a first-order perturbation calculation, where the energy levels are unaffected but mixing occurs between neighboring energy levels.

The wavefunctions of the lowest two states are expressed by

$$\psi_1(x) = N \left( \cos \frac{\pi x}{2l} - \lambda \sin \frac{\pi x}{l} \right); \psi_2(x) = N' \left( \sin \frac{\pi x}{l} + \lambda \cos \frac{\pi x}{2l} \right); \lambda = 32bl/(7\pi^2) \quad (2)$$

where  $V(x) = bx$  and the infinitely high walls are located at  $x = \pm l$  (see Fig. 6). These two wavefunctions are very similar to those of the bonding and anti-bonding orbitals of heteropolar diatomic molecules or those of the  $\pi$ -molecular orbitals of acetone.

As the parameter  $\lambda$  or  $b$  becomes larger, the wavefunction for the lowest state expressed by eqn. (2) will have an unexpected node, revealing the limit of the first-order perturbation calculation in this case.

#### *A champagne bottle-shaped well potential*

A symmetrical double minimum potential is also approximated by a sine (cosine) curve as shown in Fig. 7, where the results of the first-order perturbation calculation are given [4]. When the central potential barrier is low, the effect of the perturbation is very small affecting only the shape of

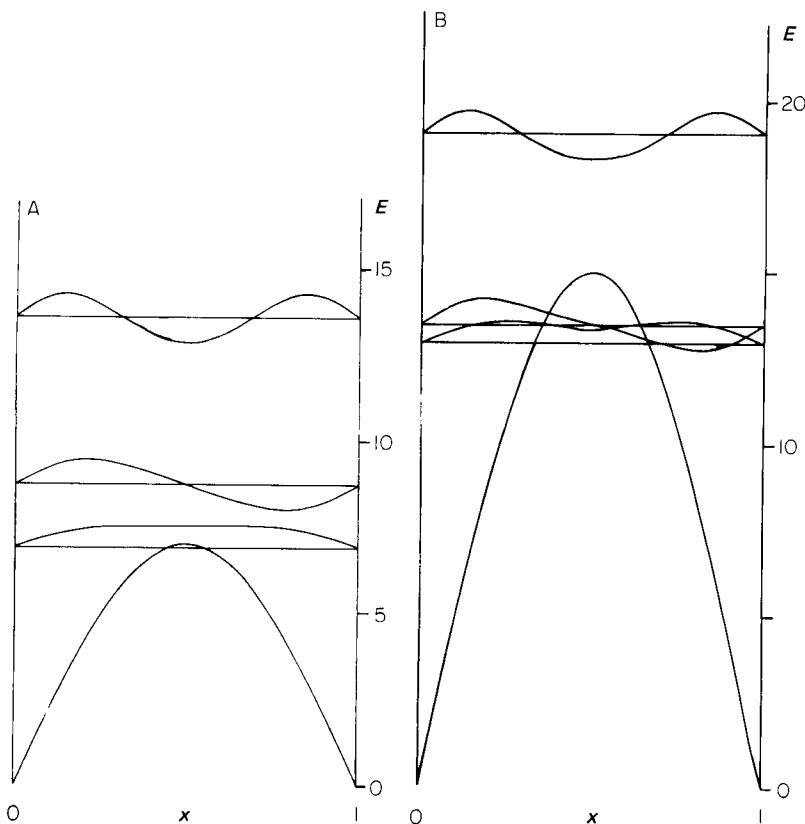


Fig. 7. Champagne bottle-shaped potentials.  $V(x) = b \sin (\pi x/a)$  with (A)  $a = 1$  and  $b = 7h^2/8m$  and (B)  $a = 1$  and  $b = 15h^2/8m$ .

the wavefunction of the lowest state. As the perturbation grows the lowest two states gradually close together and the tunnelling effect also appears as in Fig. 7B. As long as the perturbation does not break the symmetry of the whole potential, the first-order perturbation calculation gives good results.

### *V-shaped potential*

For a V-shaped potential,  $V(x) = ax$ , the solution of eqn. (1) is expressed in terms of the Airy function:

$$Ai(s) = \begin{cases} \frac{1}{\pi} \left(\frac{s}{3}\right)^{1/2} K_{1/3}\left(\frac{2}{3}s^{3/2}\right) & \text{for } s > 0 \\ \frac{1}{3} |s|^{1/2} \left\{ J_{1/3}\left(\frac{2}{3}|s|^{1/2}\right) + J_{-1/3}\left(\frac{2}{3}|s|^{1/2}\right) \right\} & \text{for } s < 0 \end{cases} \quad (3)$$

where  $K$  and  $J$  are modified Hankel and Bessel functions, respectively [2]. In order to avoid divergency, special precautions are necessary for evaluating these auxiliary functions at small and large arguments. (The library program of  $Ai$  which was recently coded into the Mathematical Subprogram Library at the Computer Center of Tokyo University was used here.) A typical result is given in Fig. 8, where it can be seen how the phase of a sine curve is distorted and the wave tails off. The spacing between neighboring levels gradually diminishes as the energy increases. After a little modification, a typical potential energy curve of a covalent diatomic molecule can be approximated by the Airy function [5].

### *Eckart potential or bell-shaped potential*

Most of the potentials introduced above are somewhat artificial in the sense that the first derivative of the potential curve becomes discontinuous in some places. Eckart solved the equation with the potential

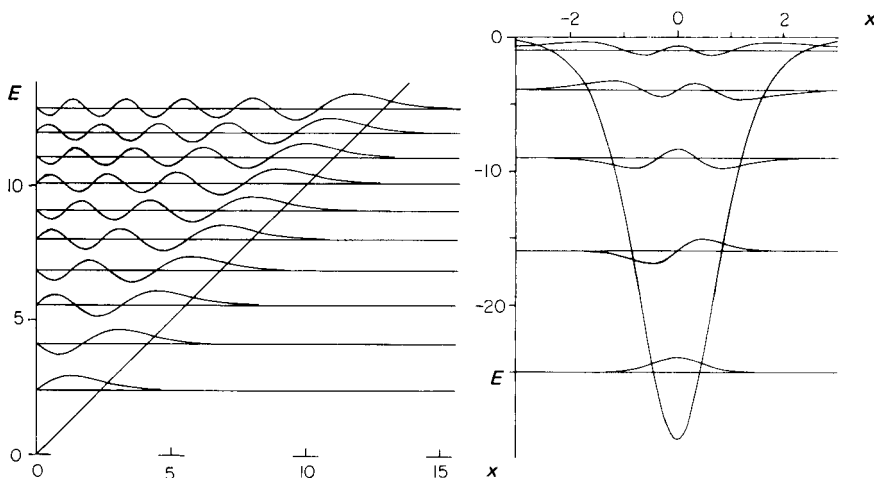


Fig. 8. A V-shaped potential.  $V(x) = ax$  (in units of  $\hbar^2/2m$ ) with  $a = 1$ .

Fig. 9. Symmetrical Eckart potential with  $\lambda = 6$  and  $\alpha = 1$  (see eqn. 5).

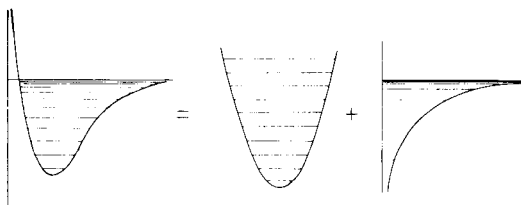


Fig. 10. Schematic figure representing the composite nature of the solutions of the Schrödinger equation. A Morse-like potential may be approximated as a combination of the potentials of a harmonic oscillator in the bonding region and of a long-range Coulombic interaction.

$$V(x) = -A\xi/(1 - \xi) - B\xi/(1 - \xi)^2; \xi = -\exp(2\pi x/l) \quad (4)$$

and obtained the result expressed by hypergeometric functions [6, 7], which, however, are generally not easy to calculate [8]. A continued fraction was found to be a powerful technique for calculating certain types of hypergeometric functions [9].

A result for a symmetrical Eckart potential

$$V(x) = -\frac{\hbar^2}{2m} \frac{\alpha^2 \lambda (\lambda - 1)}{\cosh^2 \alpha x} \quad (5)$$

is shown in Fig. 9. The parameter  $\lambda$  determines the depth and therefore the number ( $[\lambda] - 1$ ) of the bound states, while  $\alpha$  adjusts the effective width of the potential well. Note that the energy levels for this bell-shaped potential are expressed simply as  $-an^2$ . However, if  $\lambda$  is fairly large, the lowest levels lie with almost the same spacings as in the case of a parabolic potential,  $V(x) = ax^2$ , which is known as the potential of a harmonic oscillator. However, the tail of this potential behaves as  $e^{-2\alpha|x|}$ . Qualitatively, this is similar to the case of one-dimensional coulombic potential for which the energy levels are expressed as  $-C/n^2$ . This means that a long-range potential can accommodate an infinite number of states, whereas only a finite number of states is allowed for a short-range exponential potential.

### Applications

The behavior of the solutions was studied for some typical potentials. If a rough idea of these relationships between the shape of the potentials and their solutions is grasped, one can reasonably guess the sketches of the energy levels and wavefunctions for a given complicated but realistic potential by combining the component solutions. A random example is given in Fig. 10.

### REFERENCES

- 1 L. I. Schiff, *Quantum Mechanics*, McGraw-Hill, New York, 1955.
- 2 S. Flügge, *Practical Quantum Mechanics*, Springer-Verlag, New York, 1974.
- 3 M. A. Morrison, T. L. Estle and N. F. Lane, *Quantum States of Atoms, Molecules, and Solids*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
- 4 G. R. Miller, *J. Chem. Educ.*, 56 (1979) 709.

- 5 R. W. Nicholls, *Chem. Phys. Lett.*, 64 (1979) 190.
- 6 C. Eckart, *Phys. Rev.*, 35 (1930) 1303.
- 7 C. Rapp, *Quantum Mechanics*, Holt, Rinehart and Winston, New York, 1971, p. 136.
- 8 M. Abramowitz and I. A. Stegun (Ed.), *Handbook of Mathematical Functions*, AMS 55, National Bureau of Standards, Washington, DC, 1964.
- 9 H. S. Wall, *Analytical Theory of Continued Fractions*, Van Nostrand, New York, 1948.

## EDUCATIONAL EQUIPMENT IN THE PERIPHERY OF A LARGE COMPUTER

KATSUNORI HIJIKATA\* and SENRO SAITO

*University of Electro-Communications, Chofu-shi, Tokyo 182 (Japan)*

(Received 23rd January 1981)

### SUMMARY

Two practical examples of educational equipment in the periphery of a large (or medium) computer are described. The management system for lessons with several video displays is illustrated by a flow chart, and the total load to the central processor is assessed. A demonstration equipment which combines a graphic display, a scan-converter and a projection TV, is shown to be valuable. Possibilities of improvement are indicated.

The Information Processing Center of this University has a multi-purpose computer which has four functions: it acts as the processor for ordinary TSS jobs and batch jobs, the host of the inter-university network, the host of the campus network to operate on-line systems, and the supporter of educational equipments. The total system is illustrated in Fig. 1. The central processor is an M-170 which has a speed of 1 MIPS (Million Instructions Per Second) and a memory of 6 Mbyte, and, being assisted by a 2300 Mbyte disc memory, is adequate for the above tasks at present.

Thirty-seven video display tubes (VDT's) are installed in classroom A where elementary information processing is taught. These terminals are also used as normal TSS units outside class hours. The demonstration equipment in classroom B comprises a graphic display, a scan-converter and a projection TV. The system should be useful for education in any field of science or technology.

### EDUCATION BY MEANS OF VIDEO DISPLAY TUBES

In classroom A, 2 VDT's are for the teacher, and 35 for students. These terminals and a line printer are connected to the central processor by coaxial cables. At present, 11 classes are conducted in this room, the number of students being between 30 and 70 (2 students may share a terminal), and are convened once a week (for 1.5 or 3 h). In addition to regular classes, students have free access to the terminals for 50 min every morning.

The class-management system is illustrated by the flow chart in Fig. 2. At the beginning of a day, an operator inputs the schedule for the day, by which students in a particular class may use the machine for a particular period of

## UEC-IPC System

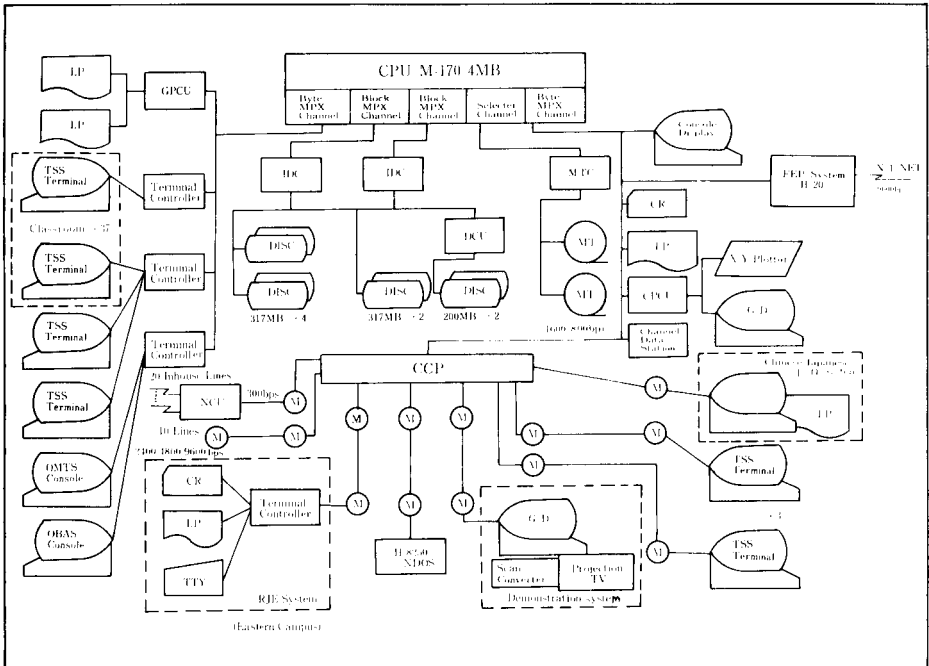


Fig. 1. The computer system.

time. One lesson is commenced by the teacher's command, `LOPEN xxxxxxxx` (teacher's ID), and this is followed by the student's input of the command, `LATTEND xxxxxxxx` (student's ID), by which his attendance is recorded in the teacher's data set. The `LSTATUS` command helps the teacher to check the record. Then the teacher begins mass education by `LDOPEN`. In the mass-education mode, the teacher communicates his messages and questions to the students' terminals. When a student answers the question, he inputs `LANSWER` followed by his answer. By `LMONITOR` plus a student's ID, the teacher can scan the student's terminal. The same command can be used for the students to see the teacher's terminal. The answers are transferred to the teacher's data set and are examined by the answer analyzing program (AAP). The mass-education mode is terminated by `LDCLOSE` and changed to the TSS mode. (Unless the teacher inputs `LDOPEN`, the terminals are in TSS mode.) Then the students start to work on exercises. In case of a FORTRAN exercise, his program may be checked and analyzed by the FORTRAN monitor program. In this mode, he can call the CAI program and learn by himself. The finished program (or answer) can be printed immediately by logging, or may be written in the student's data set by `REPORT`. The teacher can collect the answers of the students in his data set. The lesson is closed by the teacher's command `LCLOSE`.

A student may keep his data set up to 200 kbyte. The charge to a student is counted by the formula: Terminal-on time (min)  $\times$  3 yen + CPU time (s)





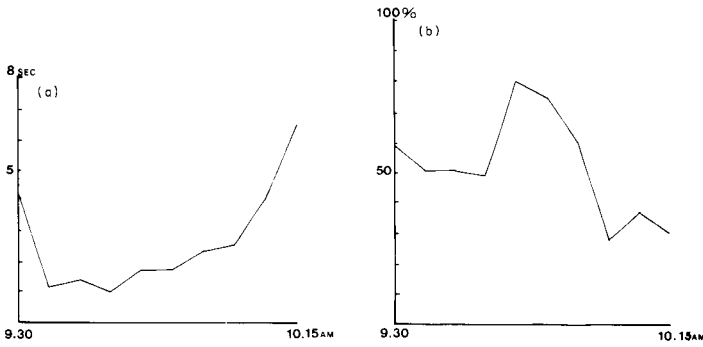


Fig. 3. Response time and percentage of CPU service to terminals. For this period of time, the VDT's are partly used by the class students, and partly by research workers. For explanation of (a) and (b), see text.

80 transactions per minute. On average, after the congestion of initial logging-on, most of the users call out their old program to be corrected. The work load in this transaction is very light and response is quick.

The percentage of CPU service given to these terminals (Fig. 3b) is initially between 60 and 80%, but when other jobs are put in about 10 a.m., it goes down to 30%; at the same time, terminal users start to run their programs, causing an inevitable increase in the response time (without complaint yet). When no class is taught, these terminals are open to ordinary users. Thus education and research work coexist peacefully at present.

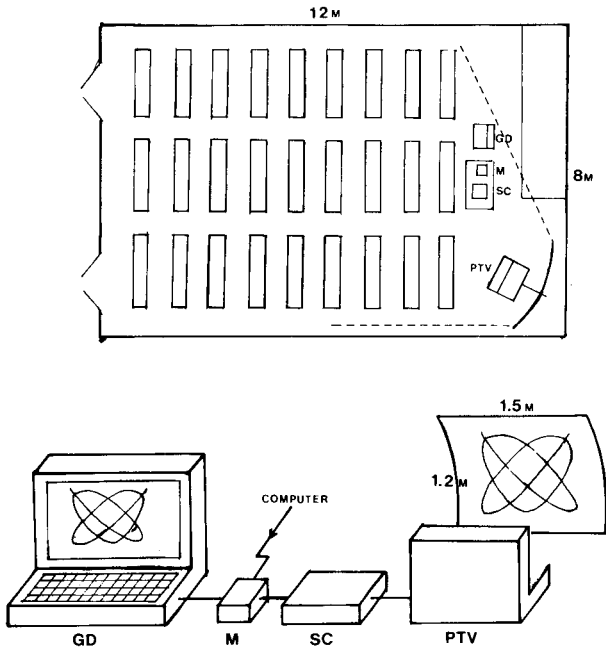


Fig. 4. Layout and components of demonstration equipment.

## DEMONSTRATION EQUIPMENT

The layout of the demonstration equipment is shown in Fig. 4. This equipment comprises a graphic display, a scan-converter and a projection TV set. The whole system is connected with the CPU by a dedicated campus telephone line. The scan-converter accepts the same signal on its Brown tube as on the graphic display, and converts it to a television signal. This instrument was obtained from Hughes Aircraft Co. Ltd., Carlsbad, Calif.

The demonstration equipment is mainly used for two purposes. First, the graphic display functions as a VDT; the content of a data set can be projected on the screen, e.g., numerical tables can be shown. Secondly, the figure on the screen can be seen during the process of drawing and its shape can be varied by putting in numerical data from the keyboard. As an example of the second case, the trajectory of a particle in a certain field of force with given initial conditions can be drawn. Figure 5 shows an interesting example involving solution of a one-dimensional Schrödinger equation, where a tentative eigenvalue is put in, producing a divergent solution, and the eigenvalue is corrected until the solution satisfies the boundary conditions.

In these examples the growth of the curve is immediately visible on the VDT, with an educational effect never attained by any other kind of teaching material. Of course, the process of drawing the curve must be suitably retarded so that its growth can be followed visually. The mesh of the graphic display used in  $4000 \times 4000$ , which is enough for this equipment to be used for more practical problems where the analytical behavior of the solution for an arbitrarily given potential is unknown. Unfortunately, the graphic display is not of a refresh type but of an accumulation type. Thus a curve can be drawn but cannot be moved, which is a great restriction.

However, it is possible to store in the machine numerical tables, formulas, etc., together with the programs for curve formation, and by correlating these a variety of attractive demonstrations can be developed.

## MAINTENANCE AND IMPROVEMENT

The educational system and equipment are the products of cooperation between the Information Processing Center and Hitachi, Ltd; the latter followed our specifications in detail, and take responsibility for the products as well as improving the software on request. This cooperation is included in the rental contract; at present a maintenance engineer is in charge of the educational system.

The VDT terminals are properly designed for TSS, where the break-in key dominates everything. There is no way to force the students' terminals over to the mass-education mode, which is an unavoidable defect. The long response time is known as a serious defect of Hitachi's TSS, and will be radically improved in the near future. A graphic display of the refresh type is very desirable but too expensive at present. Preferably, the whole demon-

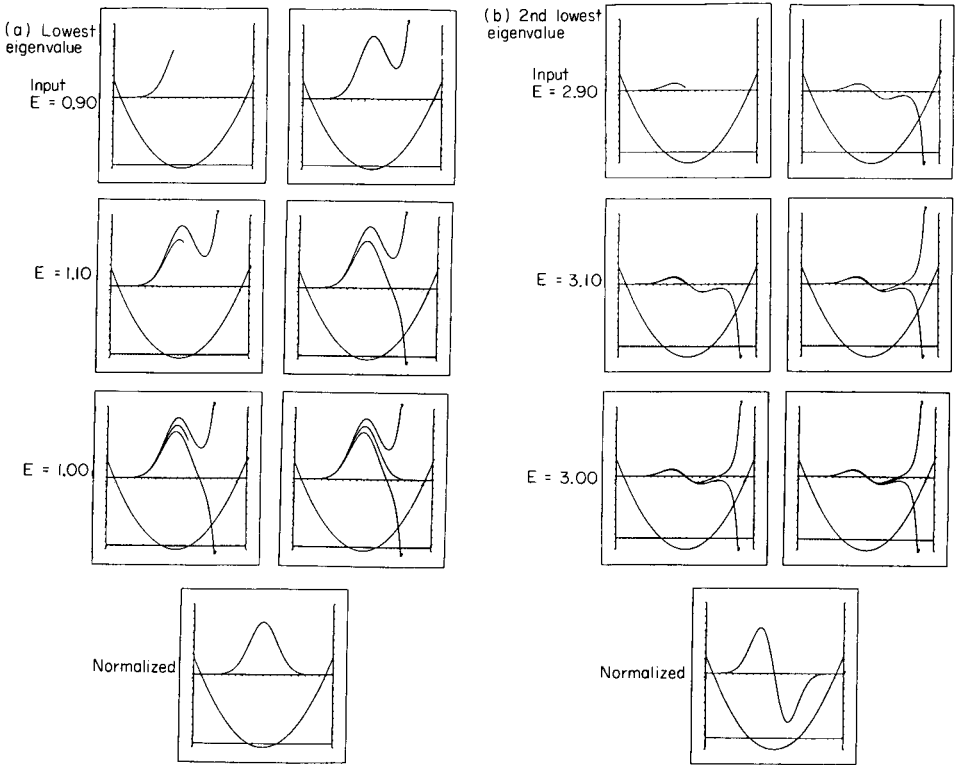


Fig. 5. Solutions of eigenvalue problem:  $d^2\psi/dx^2 + (E - x^2)\psi = 0$ . Boundary conditions:  $\psi = 0$  at  $x = \pm\infty$ . (It takes 10 s to draw one curve.)

stration equipment should be made portable, so that it can work by plugging in an active terminal in any classroom.

We thank FHL and Hitachi Ltd. and the members of the Information Processing Center for their cooperation in the establishment and the maintenance of the system described. Also we acknowledge the assistance of Mr. Yoichi Iida, student in the Department of Engineering Physics, who prepared the demonstration program for the eigenvalue problem.

Short Communication

---

PRELIMINARY STUDY OF A TRACE CHARACTERIZATION  
INFORMATION SYSTEM

NOBUYUKI TANAKA

*Department of Chemistry, Faculty of Science, Tohoku University, Sendai 980 (Japan)*

TAKAKO MATSUDA and MASATO SAKATA

*Computer Center, Tohoku University, Sendai 980 (Japan)*

(Received 23rd January 1981)

*Summary.* A design study of the trace characterization information system TRACIS, which has been partly implemented, is presented. TRACIS provides information on methods for trace characterization along with relevant fundamental data (e.g., physical constants and spectra) as required. The COOD data-base management system is considered to be most appropriate for TRACIS and is explained with examples. A preliminary design of TRACIS is presented.

Trace characterization is one of the most important and interesting subjects in analytical chemistry. For trace characterization, well-organized information is needed for selection of the most appropriate method as well as accurate information on the properties of materials and characteristics of reactions. Such information should be available on demand. In order to fulfil such requirements, a trace characterization information system (TRACIS) is being developed by combining computer data bases and a computer network. A preliminary study on TRACIS, based on studies carried out since 1979, is reported below.

*Data-base management system*

As TRACIS provides, on demand, information on methods for trace characterization and on fundamental data such as physical constants and spectra, a data-base management system and a computer network are essential. Many data bases are available to research workers, most for literature data, but some for numerical data. These data bases are usually created and managed by computer specialists using either general-purpose or specially-designed management systems.

In the further development of TRACIS, it is intended that data bases on particular subjects will be created and managed by research workers specializing in those subjects. General-purpose data-base management systems are comprehensive and versatile but too complex in operation for use by research chemists. Accordingly, the data-base management system

COOD [1-3], which has been developed as a user-oriented system for on-line data storage and retrieval, is used in addition to other systems.

COOD provides four types of data-base codes as well as manipulation commands. The conversational data manipulation code (CML) is self-contained and offers an easy means of data manipulation, whereas the data-manipulation code (DML) has functions for connecting the data base with FORTRAN programs. In addition, data-base operation commands are provided. The principal CM commands are given in Table 1. The COOD data model is an aggregate of tables which are composed of items, and is described in data description code (DDL). File definitions such as the capacity and permission for use of the data base or table are described by simple codes (FDL). Definitions of the data base ISES (ion-selective electrodes and sensors) written in DDL and FDL codes are given in Table 2 as examples. The data base can be created easily by using DD, FD and the command DFC. The storage of data may be done in three different ways: input of data in the

TABLE 1

## Principal CML commands

---

```

.USE [< umc > /] < data-base name > /

  {< table name > [= < tname > ]
    [ { {< item name > [= < iname > ] , ... } } ] , ... ;
    [ { ALL-ITEMS } ] ] ]

.SELECT [*n] {< tname > } , ... [ { {< iname > } , ... } ]
  [ { ALL-ITEMS } ] ]

  [ TO { < tname >
    [ [ TERMINAL- ] < file name > ] } ] [WHEN (< condition >)];

.ASK [*n] {< tname > } , ... [ { {< iname > } , ... } ]
  [ { ALL-ITEMS } ] ]

  [WHEN (< condition >)];

.STORE [*n] { NEW } < tname > [ { {< iname > } , ... } ]
  [ { OLD } ] [ { ALL-ITEMS } ] ]

  [FROM < file name >];

.CHANGE [*n] < tname > [ { {< iname > } , ... } ]
  [ { ALL-ITEMS } ] ]

  [FROM < file name > ] [WHEN (< condition >)];

```

---

TABLE 2

Definition of data base ISES written in data description DDL and file description (FDL) codings

---

DDL;

DATABASE ISES : Ion-selective electrode and sensors;

TABLE LISES : Literature on ion-selective electrodes and sensors;

NO	(I4)	UNIQUE	: Serial number;
AUTHOR(7)	(A30)		: Author(s);
TITLE(3)	(A78)		: Title;
JOURNAL	(A78)		: Journal;
YEAR	(I4)		: Year of publication;
ADDRESS(2)	(A78)		: Address of author to communicate;
KEYWORD(5)	(A50)		: Keyword(s);
DATANO	(A78)		: Name and number of data table;
REMARKS(2)	(A78)		: Remarks;
COMPILER	(A78)		: Compiler(s);

TABLE INISES : Investigations on or with ion-selective electrodes and sensors;

NO	(I4)	UNIQUE	: Serial number;
PURPOSE	(A78)		: Purpose of investigation;
ELEMENT(7)	(A30)		: Element(s) or compound(s);
MATRIX(2)	(A78)		: Matrix of element(s) or compound(s);
METHOD	(A78)		: Method or technique;
ELECT	(A50)		: Electrode or sensor;
TYPELECT	(A78)		: Type of electrode or sensor;
NAMELECT(2)	(A78)		: Name of electrode or sensor;
TITRANT	(A78)		: Titrant for titration;
RANGE	(A50)		: Range available;
RANGEN(range)	(J10)		: Range available after normalization (M);
INTERFER(5)	(A30)		: Interference;
PRETREAT	(A78)		: Pretreatment;
LITNO	(A78)		: Name and number of literature table;
REMARKS(3)	(A78)		: Remarks;
COMPILER	(A78)		: Compiler(s);

END-DDL;

FDL;

```
DATABASE ISES;
  PERMISSION READ;
TABLE LISES;
  MAX 100;
TABLE INISES;
  MAX 100;
```

---

END-FDL;

question and answer mode, transfer of data from the sequential file prepared preliminarily, and storage of data with the FORTRAN program.

Data are also retrieved in various ways. Data retrieved may be displayed at a terminal, output on the user's file, or transferred to other tables. Table 3 shows an example of data retrieval displayed at a terminal. One of the advantages of COOD is the retrieval of data from several tables in different data bases. An example is given in Table 4, where four tables from four different data bases are used for retrieval. The output indicates that five data are found. In the further procedure, only items of the serial number in each table are displayed, although all items can be obtained if so requested.

TABLE 3

An example of data retrieval

SYSTEM ?COOD

```

COOD-PROCESS ... ?CML
?USE ISES/INISES, LISES;
EXPLAIN ITEMS OF INISES, YES OR NO? →
EXPLAIN ITEMS OF LISES, YES OR NO? →
?SELECT INISES WHEN (RANGEN<=1E - 6 & ELEMENT="ZINC");
*** END OF TABLE

```

```

*** 1 DATA FOUND.
OUTPUT DATA, YES OR NO ?Y

```

```

*TABLE INISES IN ISES
DISPLAY, NAME(N) OR EXPLANATION(E) ? →

```

```

NO          : 3
PURPOSE    : Evaluation of electrode
ELEMENT    : Copper(II)
              Zinc
              Mercury(II)
              Lead(II)
MATRIX     : Test solution
METHOD     : Semi-automatic potentiometric titration
ELECT      : Ion-selective electrode
TYPELECT   : Perchlorate electrode
NAMELECT   : Orion 92-81 Perchlorate Electrode
TITRANT    : Manganese(II)
RANGE      : 1.0E - 6 M to 7.0E - 4 M
RANGEN     : 1.0E - 6 to 7.0E - 4
INTERFER   :
PRETREAT   : Without
LITNO      : LISES-2
REMARKS    :
COMPILER   : Okazaki S., Tanaka N.

```

```

?SELECT*1 LISES(AUTHOR,TITLE,JOURNAL,YEAR) WHEN(NO=2);
DISPLAY, NAME(N) OR EXPLANATION(E) ? →

```

```

AUTHOR     : Hadjiioannou T. P.
              Koupparis M. A.
              Efstathiou C. E.
TITLE      : Evaluation of a perchlorate-selective electrode for catalytic titrations involving
              periodate indicator reactions
JOURNAL    : Anal. Chim. Acta, vol. 88, p. 281-287
YEAR       : 1977

```

?

### Computer network

An experiment to link two computer systems at different large-scale computer centers was started almost ten years ago by a research group at Tokyo University and Kyoto University. The two different computer systems were linked with the N-1 protocol developed by the research group [4]. The

TABLE 4

An example of data retrieval from four different data bases

---

```

COOD-PROCESS ...?CML
?USE user-id.4/REFAY/KIPERIS=AY;
EXPLAIN ITEMS OF KIPERIS, YES OR NO ? →
?USE user-id.5/REFMS/REMECS=MS;
EXPLAIN ITEMS OF REMECS, YES OR NO ? →
?USE user-id.1/REFEK/REMECS=EK;
EXPLAIN ITEMS OF REMECS, YES OR NO ? →
?USE user-id.6/REFTY/DICISIS=TY;
EXPLAIN ITEMS OF DICISIS, YES OR NO ? →
?SELECT AY, MS, TY, EK(NO)
MORE?WHEN((CONHYD>=1.26E - 5 & CONHYD<=2E - 5), (PH>=4.7 & PH<=4.9));
*** END OF TABLE
***   ON DATABASE   user-id.4/REFAY   /KIPERIS
*** END OF TABLE
***   ON DATABASE   user-id.5/REFMS   /REMECS
*** END OF TABLE
***   ON DATABASE   user-id.6/REFTY   /DICISIS
*** END OF TABLE
***   ON DATABASE   user-id.1/REFEK   /REMECS

***   5 DATA FOUND.
OUTPUT DATA, YES OR NO ?Y

*TABLE KIPERIS IN REFAY
DISPLAY, NAME(N) OR EXPLANATION(E) ? →

NO   :   1

NO   :   2

NO   :   3

*TABLE REMECS IN REFMS
DISPLAY, NAME(N) OR EXPLANATION(E) ? →

NO   :   1

*TABLE REMECS IN REFEK
DISPLAY, NAME(N) OR EXPLANATION(E) ? →

NO   :   7

?

```

---

results indicated the possibility of using computer networks in information system to be developed.

A new communication system called DDX (digital data exchange) started its service in July 1980. The Computer Center of Tohoku University whose computer system is different from those of both Tokyo and Kyoto Universities has also started experiments with the computer network, with promising results.



### Design of TRACIS

The basic design of TRACIS is shown in Fig. 1. In order to control input information, the CIPS (central information processing system) is considered, in which the computer network is to be included. There are five categories of data bases: (i) data on analytical methods, (ii) physical constants and spectral data, (iii) data on materials, (iv) application programs, and (v) information on research projects. Those data bases are generally located at the research group which creates and manages them, and so may be located at several computer centers. At the moment, the data bases are located at the Computer Center of Tohoku University, because the computer network is not operational. Data bases created at other centers are transferred to and managed at the Computer Center of Tohoku University.

The system structure of TRACIS (Fig. 2) shows that TRACIS contains various data-base management systems at the formation and management site and at the retrieval site. The present TRACIS does not include the computer network in the CIPS but will later be established as a multi-center multi-DBM system. A procedure for access of TRACIS and registered data bases is presented in Table 5.

The authors express their sincere gratitude to all members of the Research Project "Information Studies on Trace Characterization" for their constant cooperation, especially Prof. Hiroshi Inose and Dr. Hisashi Yasunaga for

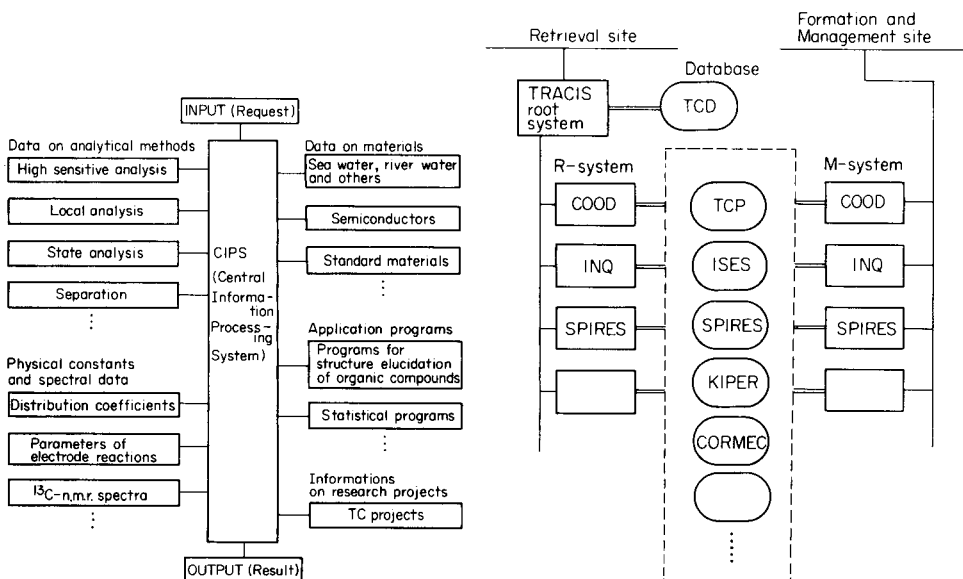


Fig. 1. Preliminary design of TRACIS.

Fig. 2. System structure of TRACIS.

TABLE 5

Procedures for access of TRACIS and of data bases registered

---

SYSTEM? TRACIS

\*\*\* TRACIS on Computer Center, Tohoku University \*\*\*

Date : 09/26/80

---- List of TC-Database ----

TCP: Projects and Investigators of Trace Characterization  
Composed of two tables,  
PROJECT (Project) and MEMBER (Member).

ISES: Ion-selective Electrodes and Sensors  
Composed of two tables,  
LISES (Literature on ion-selective electrodes and sensors)  
INISES (Investigations on or with ion selective electrodes and sensors).

(Omitted)

Do you need more information, Yes or No ?Y

Which DB ?TCP

Served by Computer Center, Tohoku University

ACOS 77/900, ACOS-6

Managed by Nobuyuki Tanaka (Department of Chemistry, Faculty of Science, Tohoku University)

Used DBMS is COOD

Directions for use:

Use CML of COOD.

?SELECT table ..... ;

Do you wish to retrieve this DB, Yes or No ?Y

Now retrieval starts. Use CML of COOD.

---- Start retrieval by COOD----

(Retrieval of data base desired)

---- End of retrieval ----

Do you need more information, Yes or No ? →

SYSTEM ?

---

their advice and cooperation. The authors also thank the Ministry of Education, Science and Culture for financial support (Project nos. 389010, 411701 and 510801).

#### REFERENCES

- 1 T. Matsuda and N. Tanaka, *Trans. Inf. Process. Soc. Jpn.*, 21 (1980) 347.
- 2 N. Tanaka (Ed.), *COOD Language Specification (2.0), and COOD Reference Manual (2.0)*, Tohoku University, Sendai, 1979.
- 3 N. Tanaka, T. Matsuda and A. Yamada, *Proc. 7th International CODATA Conference, Kyoto, 1980*, in press.
- 4 H. Inose, *Research on Scientific Information System in Japan*, 1980, p. 25.

Short Communication

---

CONSTRUCTION OF A DATA BASE FOR COBALT-59 NUCLEAR  
MAGNETIC RESONANCE SPECTROMETRY

AKIRA YAMASAKI

*Laboratory of Applied Chemistry, The University of Electro-Communications, Chofu,  
Tokyo 182 (Japan)*

(Received 23rd January 1981)

*Summary.* The TOOL-IR PDB and COOD systems are compared for the construction of data bases for  $^{59}\text{Co}$ -n.m.r. bibliographic and spectral data. The spectral data used are the chemical shifts from several different standards, and the line widths and coupling constants (if present). The PDB system is effective for storage and retrieval of bibliographic data, but the COOD system is better for the retrieval of spectral data, and for combination of data files on literature and chemical shifts.

Recently, there has been increased demand for computer-readable data bases that can be readily constructed from personal collections of literature, numerical and graphical data. For this purpose, it is usually necessary to use data-base management systems suitable for non-specialists in computer operation. Many computer systems have their own data-base management systems, but such systems are designed for much larger data bases than the usual personal collections and they need to be operated by professional computer specialists.

To overcome such difficulties, two data-base management systems were constructed recently, namely, TOOL-IR PDB (University of Tokyo) [1] and COOD (Tohoku University) [2]. The construction of miniature data bases by means of these two systems from collections of cobalt-59 n.m.r. data from the literature and from spectra is reported in this communication.

*Cobalt-59 nuclear magnetic resonance*

In 1951, the first  $^{59}\text{Co}$ -n.m.r. spectra of several cobalt(III) complexes were reported [3] and it was pointed out that there are very large chemical shift differences ranging around several percent. Since then, there have been about seventy reports on  $^{59}\text{Co}$ -n.m.r. observations and theoretical treatment [4]. The correlation of chemical shift and crystal-field splitting of 3d orbitals (Griffith—Orgel plot [5]) is important in explaining the general heteroatom n.m.r. chemical shifts.

The  $^{59}\text{Co}$ -n.m.r. spectral line widths are determined mainly by quadrupolar relaxation, which depends on the electric field gradients at the central cobalt nucleus in the complexes [6, 7]. The line widths are therefore greatly affected by the configuration of six ligand atoms, and the chemical shift

and line width data from  $^{59}\text{Co}$ -n.m.r. should provide very useful information for characterization of many diamagnetic cobalt complexes. Until recently, the total chemical shift data of cobalt-59 were around 1500 except for alloys, metals, magnetic oxides, and some other solid samples.

The precision of chemical shift measurements has increased since the introduction of the Fourier-transform technique. The large chemical shifts of  $^{59}\text{Co}$ -n.m.r. spectra are valuable in characterizing and distinguishing complex isomers such as the four isomers of tris(*l*-propylenediamine)cobalt(III) complex cations [8] which are difficult to recognize from electronic spectra, because the difference in the *d*-*d* transitions is only several Ångström.

The previous summary of  $^{59}\text{Co}$ -n.m.r. spectral data [4] contains some disastrous errors because of wrong identification of complexes and miscalculation of chemical shifts. Moreover, the large chemical shifts make it necessary to use different standard substances, and the chemical shift data with such different standards cannot be compared simply by subtraction.

Accordingly, in the preparation of  $^{59}\text{Co}$ -n.m.r. data bases, the chemical shift data should be retrievable from the complex name (systematic and conventional names), the complexes should be retrievable from the chemical shift data measured from different standards, and the original references should be obtained along with the n.m.r. data.

#### *The TOOL-IR PDB system*

TOOL-IR is the name of the information retrieval system of the Computer Centre, University of Tokyo [9]. This system can now be used for retrieval from the following data bases: (1) CA SEARCH; (2) XDC; (3) INSPEC Part-C; (4) COMPENDEX; and (5) Ecology and Environment.

The PDB (personal data base) system was constructed as a subsystem of this TOOL-IR for the construction of, and retrieval from, small literature data bases [1]. Data bases for n.m.r. literature [10], powder engineering [11] and solvent extraction literature [12] are now available. This system is designed for operation by non-specialists and retrieval is almost the same as in the TOOL-IR system. Examples of input data for literature and  $^{59}\text{Co}$ -n.m.r. chemical shifts are shown in Table 1. The numerical data (chemical shift, line widths, etc.) must be placed in the keyword fields. To ease retrieval, the notation of different standards should be added to the chemical shift value as follows

CS: chemical shift from hexacyanocobaltate(III)

ES: chemical shift from tris(ethylenediamine)cobalt(III)

NS: chemical shift from hexanitrocobaltate(III)

LS: chemical shift from hexaamminecobalt(III) (luteo complex)

AS: chemical shift from tris(acetylacetonato)cobalt(III) (chloroform solution)

The retrieval output is shown in Table 2.

#### *The COOD system*

The FORTRAN data-base management system COOD of the Computer Center, Tohoku University was designed by Matsuda and Tanaka [2]. This

TABLE 1

Examples of PDB input<sup>a</sup>

## Literature File Input

#00023.02: "Cobalt-59 nuclear magnetic resonance studies of stereoisomerism in tris-(bidentate)cobalt(III) complexes.", Ann Johnson and Grover W. Everett-Jr, Inorg. Chem. (INOCAJ), 12(12), 2801-5(1973).

@CoO6, CoN6, stereoisomerism@

@REFN0023@@

## Chemical Shift Data File Input

#001356.02: "Hexakis(trimethylphosphite)cobalt(III) ion in acetonitrile.", A. Yamasaki, T. Aoyama, S. Fujiwara and K. Nakamura, Bull. Chem. Soc. Japan (BCSJA8), 51(2) 643-4(1978).

@[Co(P(OCH3)3)6](3+), [CoP6] REG#; 66083-03-2 REFN0057

@CS000304 ES007700\* NS007800\* LS008400\* AS012800\*

@(1)J(Co59-P31) 414Hz, septet

@LW000340@

<sup>a</sup>An asterisk means that the value is an estimated one.

system is also designed for use by non-specialists and to link other computer programs such as statistical calculations and graph drawing. This COOD system is used here only for storage and retrieval of <sup>59</sup>Co-n.m.r. literature and chemical shift data. The system produces the user's own data tables which are appropriate data elements for data collections prior to storage. The contents of tables for literature and chemical shift data files are illustrated in Table 3 and the retrieval output is given in Table 4.

In this COOD system, different data-base files can be used together during retrieval if the appropriate assembling commands are applied [2]. Therefore, the data bases can be constructed separately in different ways, if the data tables are exactly the same.

*Discussion*

The construction of literature data bases on <sup>59</sup>Co-n.m.r. seems much easier in the PDB system than in the COOD system because the original purpose of PDB was to store and retrieve users' own literature collections. However, numerical data such as chemical shifts are less easily introduced into the data base, and retrieval needs some artificial techniques because of the character-dependent system. Construction of a data base by the conversational mode in the COOD system is somewhat time-consuming, but construction of a data base containing characters and numerical values in the same records is much easier than in the PDB system. The COOD system is therefore recommended when the originals contain both literature and numerical data; the PDB system is more versatile for literature collections.

## TABLE 2

## Examples of PDB output

//PDBSG DATA, SYMLIB (=Data File Retrieval)  
(abbreviated)

TYPE IN COMMAND

@T HEXAKIS (TRIMETHYLPHOSPHITE)COBALT (Multi-keyword Search)

T HEXAKIS(TRIMETHYLPHOSPHITE)COBALT  
RETRIEVAL KEYS: HEXAKIS/TRIMETHYL/COBALT/ETIHPSOH

K = "HEXAKIS"

K = "TRIMETHY"

K = "COBALT"

K = "ETIHPSOH"

DOCUMENT SET # 1 CREATED

SET CONTAINS 1 DOCUMENT

TYPE IN COMMAND

@DISPLAY M.D

DISPLAY M.D

DOCUMENT SET # 1 DISPLAYED

SET CONTAINS 1 DOCUMENT

-----  
SER#: 001356 TYPE: 02  
AUTH: Yamasaki, A., Aoyama, T., Fujiwara, S., Nakamura, K.  
TITL: Hexakis-(trimethylphosphite)cobalt(III) ion in acetonitrile  
HTTL: Bull. Chem. Soc. Japan CODN: BCSJA8  
VOLN: 51 ISSU: 2 PAGE: 643-4  
DPBL: 1978  
CIDX: [Co(P(OCH3)3)6](3+) [CoP6] REGNO: 66083-03-2 REFN00057  
FIDX: CS000304 ES007700\* NS007800\* LS008400\* AS012800\*  
SCOD: (1)J(Co59-P31) 414Hz. septet UCOD: LW000340

//PDBSG LITERAT, SYMLIB  
(abbreviated)

-----  
label for standards  
chemical shift  
-----

TYPE IN COMMAND

@SEARCH REFN0057

SEARCH REFN0057

K = "REFN0057"

DOCUMENT SET # 20 CREATED

SET CONTAINS 1 DOCUMENT

-----  
CS [Co(CN)6]<sup>3-</sup>  
ES [Co(en)<sub>3</sub>]<sup>3+</sup>  
NS [Co(NO<sub>2</sub>)<sub>6</sub>]<sup>3-</sup>  
LS [Co(NH<sub>3</sub>)<sub>6</sub>]<sup>3+</sup>  
AS [Co(acac)<sub>3</sub>]  
-----

-----  
SER#: 000057 TYPE: 02  
AUTH: Yamasaki, A., Aoyama, T., Fujiwara, S., Nakamura, K.  
TITL: Nuclear magnetic resonance of cobalt complexes. Cobalt-59 nuclear magnetic  
resonance spectrum of hexakis(trimethylphosphite)cobalt(III) complex.  
HTTL: Bull. Chem. Soc. Japan CODN: BCSJA8  
VOLN: 51 ISSU: 2 PAGE: 643-4  
DPBL: 1978  
CIDX: CoP6  
FIDX: REGNO-66083-03-2  
SCOD: CASNO 88.20. 143889z UCOD: J(Co-P) 414Hz  
-----

TABLE 3

COOD data tables and contents in the cobalt-59 literature (LIT) and chemical shift (XXX) data bases

DDL;		DDL;	
INSERT DATA PDF;		INSERT DATABASE PDF;	
TABLE LIT;		TABLE XXX;	
NO (15) UNIQUE: Identification No.		NO (15) UNIQUE: Identification Number	
AUTHOR(10) (A30): Author(s)		COMPLEX (A70): Name of Complex	
TITLE(5) (A70): Title		FORMULA (A70): Formula of Complex	
JNL (A70): Journal		CSHIFT (J10) : Co-59 Chem. Shift from [Co(CN)6]	
CODEN (A6) : Coden		ESHIFT (J10) : Co-59 Chem. Shift from [Co(en)3]	
YEAR (14) : Year		NSHIFT (J10) : Co-59 Chem. Shift from [Co(NO2)6]	
REMARK (A70): Remark		LSHIFT (J10) : Co-59 Chem. Shift from [Co(NH3)6]	
TYPE (A70): Ligand Atom		ASHIFT (J10) : Co-59 Chem. Shift from [Co(acac)3]	
	Combination	LWIDTH (A50): Line widths	
CASNO (A20): Chem. Abstr. Cit. No.		FREQ (A20): Reson. Freq.	
KEY(2) (A50): Keywords		REFNO (14) : Ref. No. in LIT	
		REGNO (A10): Registration No. of Complex	
		LCOMP (A15): Ligand Atom Composition	
		COMMENT(3) (A70): Comments.	

TABLE 4

Retrieval from the COOD system

```

?USE PDF/LIT = Y;
EXPLAIN ITEMS OF LIT, YES OR NO?
?USE PDF/XXX = Z;
EXPLAIN ITEMS OF XXX, YES OR NO?
?SELECT Z(CSHIFT, REFNO) WHEN (FORMULA = '[Co(en)3] (3+)');
***END OF TABLE
***5 DATA FOUND
OUTPUT DATA, YES OR NO? YES
*TABLE XXX          IN PDF
DISPLAY, NAME(N)   OR EXPLANATION(E)?
CSHIFT : 7300(-)
REFNO  : 1
CSHIFT : 7180(-)
REFNO  : 3
CSHIFT : 7380(-)
REFNO  : 7
CSHIFT : 7120(-)
REFNO  : 10
CSHIFT : 7270(-)
REFNO  : 14
?SELECT Y WHEN (NO = 1);
***END OF TABLE
OUTPUT DATA, YES OR NO? YES
***1 DATA FOUND
*TABLE LIT IN PDR
DISPLAY, NAME(N) OR EXPLANATION(E)?
NO      : 1
AUTHOR  : Proctor, W. G.
        : Yu, F. C.
TITLE   : On the nuclear magnetic moments of several stable isotopes
JNL     : Phys. Rev., 81, 20-30 (1951)
CODEN   : PHRVAO
YEAR    : 1951
REMARK  : The first cobalt-59 NMR data
TYPE    : CoC6, CoN6
CASNO   :
KEY     :

```

The author expresses his sincere gratitude to Professor Nobuyuki Tanaka and Miss Takako Matsuda, Tohoku University, for the construction and use of the COOD System, and to Professor Takeo Yamamoto and Professor Masamitsu Negishi for the construction of the TOOL-IR PDB system and helpful discussions. A Scientific Research Grant-in-Aid from the Ministry of Education, Japan, is gratefully acknowledged.

## REFERENCES

- 1 M. Negishi and T. Yamamoto, TOOL-IR Documents No. 9 (1977).
- 2 T. Matsuda and N. Tanaka, *Joho Shori Gakkai Ronbun Shi*, 21 (1980) 347.
- 3 W. G. Proctor and F. C. Yu, *Phys. Rev.*, 81 (1951) 20.
- 4 A. Yamasaki, *Rep. Univ. Electr.-Commun.*, 27 (1977), 291; 29 (1978) 69.
- 5 J. S. Griffith and L. E. Orgel, *Trans. Faraday Soc.*, 53 (1957) 601.
- 6 H. Hartmann and H. Sillescu, *Theor. Chim. Acta*, 2 (1964) 371.
- 7 A. Yamasaki, F. Yajima and S. Fujiwara, *Inorg. Chim. Acta*, 2 (1968) 39.
- 8 Y. Koike, F. Yajima, A. Yamasaki and S. Fujiwara, *Chem. Lett.*, (1974) 177.
- 9 T. Yamamoto, M. Negishi, M. Ushimaru, Y. Tozawa, K. Okabe and S. Fujiwara, *Proc. 2nd U.S.A.-Japan Computer Conf.*, Tokyo, 1975, p. 159.
- 10 A. Yamasaki, K. Kurokawa, B. Nagao, M. Negishi, T. Yamamoto and S. Fujiwara, *Rept. Univ. Electr.-Commun.*, 30 (1979) 89.
- 11 Y. Yanagisawa and T. Inoue, TOOL-IR Documents No. 10 (1978) 27.
- 12 A. Yamasaki, M. Abe, E. Yamagishi, T. Sekine and Y. Hasegawa, *Rep. Univ. Electr.-Commun.*, 31 (1981) 249.



Short Communication

---

MANAGEMENT AND QUERYING OF MORPHOLOGICAL,  
PHYSIOLOGICAL, BIOCHEMICAL AND CHROMATOGRAPHIC DATA  
DESCRIBING MICROBIAL STRAINS

MICAH I. KRICHEVSKY

*Microbial Systematics Section, National Institute of Dental Research, National Institutes of Health, Bethesda, Maryland 20205 (U.S.A.)*

(Received 23rd January 1981)

SUMMARY

An interactive information system, MICRO-IS, is described. It comprises morphological, physiological, biochemical, and chromatographic data describing microbial strains. A wide variety of retrieval, management, and editing functions can be invoked with easy-to-learn, user-oriented commands.

Most attributes of microbial strains are coded as binary values, with a lesser number coded as multistate or numerical values. Easily learned methods have been developed for coding, entering, file handling, querying, and analyzing a wide variety of data describing strains of microbes. The system is installed in the U.S.A. on the IBM 370 complex at the National Institutes of Health and in Japan under the management of the Office of Life Science Promotion, Institute of Physical and Chemical Research. While this microbial information system (MICRO-IS) was developed for research and regulatory work, the logic of the system is quite suitable for student use.

In research, routine and educational microbiological work, the operations of isolation, characterization, taxonomy, identification and general data management are needed. Often, the only difference is the scale of the operations to be performed, but students often find even the lessened scale a problem because of inexperience. Computer aid in performing the functions listed in the first paragraph can be just as useful to students.

Whether or not the microbiologist is familiar with computer science and technology (and most are not), the computer system should function as a simple tool and not become a distraction from the basic task of interacting with one's data. The MICRO-IS system is therefore user-oriented. The system is simple to learn and relearn. Ease of relearning is especially important because use of particular system elements may be sporadic. Ease of learning and use is a basic design parameter of the whole MICRO-IS and should be borne in mind throughout the following discussion.

### *Description of the system*

The coding method used is an extension of that first described by Rogosa et al. [1]. The method is comprehensive and flexible in that it allows for coding of about 10,000 features of bacteria, protozoa [2], and selected groups of fungi (including yeasts). Addition of features of algae is a current project. In order to assure uniformity of coding, a unique six-digit number is assigned to each feature. The numbers are used to label all data on entry into the computer and all subsequent operations involving specific features of strains. The first of four ways of entering data into MICRO-IS is through use of specially designed coding sheets [1, 3]. Information on nomenclature, strain history, and general comments are entered in alphanumeric format in specific areas of a series of code sheets. All other data (binary, multistate, and numerical) are entered on one kind of code sheet. For coding purposes, each state of a multistate feature is entered as a single binary feature. Binary and numerical information can be entered in any desired mixture in columns on the sheets. The meaning of each column of data is conveyed to the computer by user-supplied headings (the 6-digit feature numbers) on each code sheet. Since each code sheet is treated as an independent entity in the computer, the user is permitted great flexibility in the order and content of data entered. Code sheets and keypunching are used for high-volume batch entry of data.

The MICRO-IS is designed as an interactive system. With this in mind, the second mode of data entry is through computer terminals. The program which supports on-line data entry also is used for editing existing data. The user can select any order and amount of data to input within a given session. The order and amount can be changed at any time within a session. Data entry formats can be stored for later recall by each user. The formats can be temporarily or permanently edited in any session. This on-line editing and entry program is most useful for low-volume, intermittent use by students, for technicians with intermittent data to enter, and for updating existing data sets.

The two entry methods described above involve manual entry of data. The next two involve machine to machine communication. Where files of microbial strain data already exist in computers, they will not be in the format used in MICRO-IS. Most existing files are stored in fixed format, i.e., each item of data always occurs in the same position in the strain description. The user takes advantage of this fact in using the translation program of the MICRO-IS. A table is constructed of correspondence between data items in the existing data file and the MICRO-IS coding conventions. One-to-one correspondence is unnecessary. The translation operation results in a data set in exactly the same format as if the data were entered through code sheets and keypunching. The most common uses of the translation of existing data files mode of entry is to make available large data bases for obtaining feature frequencies or to allow comparison of data from a variety of laboratories.

Currently under development is the computer-aided data conditioning and entry of gas-liquid chromatography (g.l.c.) data on microbial culture fluids

into the MICRO-IS. The chromatographs are interfaced with a minicomputer dedicated to calculating area and time. The g.l.c. peak areas and retention times are transmitted daily to a large computer for further processing. The first step is the normalization of each run. Each chromatograph is periodically standardized with a homologous series of straight-chain saturated derivatives of the same kind to be used for samples. A series of coefficients for cubic equations is established for each chromatographic instrument by the technique of "spline fitting". Interpolation of actual times found yields a straight-chain saturated carbon equivalence number (SSCEN). By this technique the same compound has the same SSCEN on any instrument. Calculation of concentration from areas is accomplished by well known techniques.

Programs have been written to manage, edit and retrieve the standardized runs. The user can retrieve any series of runs, place them into a temporary file, and perform algebraic addition and subtraction of runs. Thus, a derived run can be one which has a background run subtracted and is the summation of replicate runs. These derived runs are used to develop binary or numerical data for entry in the MICRO-IS. At this point, each peak may be treated as a binary or numerical item as with any other feature descriptive of a strain.

No matter which combination of the four routes of data entry is used, each strain record is stored as part of a highly compressed, searchable data set. All other programs in the MICRO-IS interact with this format. With the single exception of the search commands of the query program, all functions of the system are utilized by responding to computer-generated prompts.

The MICRO-IS has two simple file management functions. The first function copies individual data sets into a common file so that they may be searched together. The second function moves strain descriptions from within one data set into a second. The movement can be of one strain at a time or all strains meeting a specific search criterion. These two functions are adequate to meet most, if not all, management needs of system users.

A problem shared by students and technicians is identification of unknown strains. The task is one of pattern matching of the features of the isolated unknown with known taxa (usually species). The microbiologist must decide if any known taxon is both a close and unique match to the unknown. The usual answers are: the organism is of a particular taxon, is atypical for that taxon, or is unidentified with respect to the list of taxa used.

Available in MICRO-IS is a program for identification by use of conditional probabilities [4]. The program analyzes a matrix of the frequency of occurrence of each feature within each taxon to find the "best fit". At present, 14 matrices covering the most frequently encountered heterotrophic bacteria are available. With the aid of numerous collaborators, we are adding and improving matrices. In this context, as g.l.c. data are accumulated, they will be incorporated into matrices. In terms of usage, the identification program is the most frequently used program in the MICRO-IS.

While the identification program is the most often used, the query and report program is the most versatile and manipulates the most data. It is the

only program in the system which requires the user to issue commands in addition to responding to prompts (i.e., questions or instructions from the computer).

The syntax (i.e., form) of the file commands having no prompts is simple. A minimum of punctuation is used because the commonest errors made by beginners and sporadic users of command languages are those involving punctuation. Only three symbols are required ( , ), and =. The parentheses are separators (i.e., delimiters) of parts of a command similar to algebraic expressions. The = indicates a string of arbitrary characters to be searched for. Any symbol not in the search string may be used as the character string delimiters (e.g., = "ABCD" or = !ABCD!). A blank space is the only other delimiter needed.

With the commands, data sets can be searched by using simple Boolean logic (+, -) for binary data, (=, >, <) for numerical data, and character string matching for alphanumeric data. Complex searches are performed by combining any of these elements through use of logical AND, logical OR, parentheses, and an allowance for similar but not identical strains. This last is accomplished by selecting the number of elements in a search that must be true (e.g., 8 out of 10).

Depending on the verb used in a search command, one of five products is generated as a result of a successful search. The verb SHOW results in specially selected items of data on each strain (found by a logical search) being displayed at the computer terminal in tabular form. Using the same concept with the verb TAB, simple frequency, range, and summation descriptions of the selected items are listed.

The verb EXTRACT invokes the copying of the strain records for the strains found in the search into a data set which is in turn a searchable subset of the original. This command enables hierarchical searches without requiring a hierarchical command structure. Any one of the subsets may be saved permanently as a data set in its own right by use of the verb SAVE.

Three other verbs are used in the command structure that are not involved in searching strain information. The simplest is END which terminates the search program. GIVE yields a list of searchable items in any data set.

The last verb, REPORT, invokes a comprehensive prompt protocol for generating tables of data complete with headings. These tables can be displayed at the terminal or saved for later printing as well as use by other analytic programs [5].

### *Conclusion*

The main uses of MICRO-IS have been in epidemiologic, ecologic, and taxonomic research as well as in monitoring the environment for potentially hazardous microbiota. However, students have used MICRO-IS in their doctoral studies in these same disciplines. With the exception of the still developing chromatography modules, students can learn to use the system completely in two to four days. Undergraduate students also could benefit from using a system of this sort.

## REFERENCES

- 1 M. Rogosa, M. I. Krichevsky and R. R. Colwell, *Int. J. Syst. Bacteriol.*, 21 (1971) 1A.
- 2 P. R. Daggett, M. I. Krichevsky, M. Rogosa, J. O. Corliss and J. P. Girolami, *J. Protozool.*, 00 (1981) 000.
- 3 F. A. Benedict, *FDA By-Lines*, 9 (1979) 223.
- 4 R. Johnson, *FDA By-Lines*, 9 (1979) 235.
- 5 C. A. Walczak, *FDA By-Lines*, 9 (1979) 251.

## Short Communication

---

### FOUR-M CALCULATIONS ON METHANOL-WATER SOLUTIONS

SUSUMU OKAZAKI, KOICHIRO NAKANISHI\* and HIDEKAZU TOUHARA

*Department of Industrial Chemistry, Kyoto University, Kyoto 606 (Japan)*

(Received 23rd January 1981)

**Summary.** The structural characteristics of methanol in aqueous solutions, on a molecular level, can be elucidated by four types of calculation: molecular orbital, multiparametric optimization of intermolecular potential function, Monte Carlo, and molecular dynamics. As a first step, the potential between water and methanol was determined by ab initio LCAO SCF molecular orbital calculations with the STO-3G basis set and subsequent multiparametric fitting. This and water-water potentials were used for Monte Carlo calculation on an aqueous methanol solution containing a 1:216 mole ratio of methanol to water. Hydration around methanol is briefly discussed.

One of the most important applications of computer calculations in the field of physical chemistry is in liquid-state experiments. In principle, statistical mechanics can describe the thermodynamic and structural properties of matter, but, in the liquid state accurate results are difficult to obtain because of complex intermolecular interactions. Computerized experiments play an essential role as they can afford various static and dynamic properties of fluids in which molecules interact with a prefixed potential. There are two methods in such computerized experiments: one is the molecular dynamics method where the trajectory of molecular motion can be obtained from numerical solution of the Newton equation of motion; the other is the Monte Carlo method which involves the generation of statistically weighted configuration.

Computerized experiments were first done with simple liquids for which hard sphere and Lennard-Jones potentials can be used. Recent development of high-speed computers has made it possible to deal with more complex liquids in time-consuming calculations and associated liquids have already been studied. In the case of associated liquids, both empirical and non-empirical pair potentials have been used. The latter type of potential which can be obtained from quantum mechanical LCAO SCF calculation should be preferable [1]. Since sophisticated molecular orbital calculation is easily accessible at present, theoretical studies on associated solutions should include the following four calculations: (1) molecular orbital (MO) calculations for various sets of dimer configuration; (2) multiparametric (MP) optimization of the intermolecular potential function by the nonlinear least-square method from the MO results; (3) Monte Carlo (MC) calculation; and (4) molecular dynamics (MD) calculation using the pair potential determined above.

An extensive research program involving the above four-M calculations on alcohol-water solutions is planned in order to elucidate the structural and thermodynamic features of these important solutions on a molecular level. As the first step of this project, the potential between methanol and water has been newly determined by LCAO SCF calculations with the STO-3G basis set and subsequent multiparametric fitting. This and the presently available water-water potentials were used in a preliminary Monte Carlo calculation on a dilute aqueous methanol solution containing one methanol to 215 water molecules. The new pair potential for the methanol-water heterodimer and some energy information from Monte Carlo results are described briefly in this communication.

### *Methanol-water pair potential*

Ab initio LCAO SCF calculation for the methanol-water dimer was done with the STO-3G basis set by using the IMSPAC QCPE GAUSSIAN 70 program. The electron correlation and configuration interaction were not considered. The configurations of water and methanol molecules are those from microwave studies in the vapor phase [2, 3]. The conformation of methanol is fixed as the staggered form. More than 500 orientations in the dimer were generated, MO calculations were made for these orientations, those showing strong repulsion were rejected, and eventually 475 orientations were adopted for MP calculation.

These MO results were fitted to the following 23-parameter semi-empirical pair potential function:

$$V(x_1, x_2) = \sum_{i,j} q_{iW}q_{jM}(1/r_{ij}) + \sum_{i,j} a_{ij}(1/r_{ij}^3) + \sum_{i,j} b_{ij}(1/r_{ij}^6) + \sum_{i,j} c_{ij}(1/r_{ij}^{12})$$

where  $q$ ,  $a$ ,  $b$  and  $c$  are the coefficients to be optimized. This potential function is based on a multi-interaction site rigid rotor model for the two molecules (Fig. 1). In addition to the positive charges on carbon and hydrogen atoms, the water and methanol molecules bear one and two pseudo negative charges, respectively, and the distance  $R$  between the oxygen atom and the

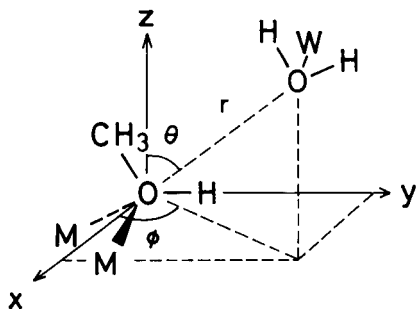


Fig. 1. Rigid rotor model for the water-methanol heterodimer.

pseudo-negative charge is also optimized. The nonlinear least-squares calculation was done by the Hooke-Jeeves method [4].

The pair potential obtained with the 23 optimized parameters can reproduce the MO results with a standard deviation of  $1.57 \text{ kJ mol}^{-1}$ . This is similar to that obtained for other water-solute potentials. Table 1 shows the parameter values.

In order to illustrate the angular dependence of the present potential, Fig. 2 shows a plot of a minimum value of the potential as a function of  $\theta$  at  $\phi = 90^\circ$  and in energetically favorable orientation. The two deep minima at  $\theta = 90^\circ$  and  $180^\circ$  correspond to two types of hydrogen bond and the higher value of the potential around  $270\text{--}360^\circ$  indicates the hydrophobic region.

#### Monte Carlo calculation for methanol-water solution

Preliminary MC calculations were done in NVT ensemble where the number of molecules,  $N$ , is 216 (216 water molecules in pure water and 1 methanol and 215 water molecules in aqueous solution), the temperature  $T$  is 298.15 K, and the volume  $V$  is that calculated from experimental density data. The MCY (Matsuoka-Clementi-Yoshimine) potential [5] with a modification by Owicki and Scheraga [6] was used to express the pair interaction between two water molecules. Generation of equilibrium configurations was done by the conventional Metropolis procedure [7]. The periodic boundary condition and minimum image convention were used. The initial state of the systems is such that the oxygen atoms occupy the Ice-I cubic lattice structure with random orientation of water as a whole. The procedure then generates  $12 \times 10^5$  configurations with an assumed prescription. The potential energy of the systems decreases rapidly with the number of configurations generated and reaches a stationary value after  $4 \times 10^5$  steps in the case of water and after  $7 \times 10^5$  steps in the case of the solution, though some further calculation would be necessary for complete convergence.

From an average over the final 5 (or 6)  $\times 10^5$  configurations, the angular dependence of the potential energy is first calculated. Two potential energy

TABLE 1

Optimized values for 23 parameters in the pair potential function between water and methanol as determined by the nonlinear least-squares method

$q_{\text{HW}}$	17.67	$q_{\text{HM}}$	9.2906	$q_{\text{CM}}$	5.7144
$a_{\text{HH}}$	27.694	$b_{\text{HH}}$	-20.826	$c_{\text{HH}}$	4473
$a_{\text{HC}}$	-19.545	$b_{\text{HC}}$	3021	$c_{\text{HC}}$	5529.989
$a_{\text{HO}}$	-0.2846	$b_{\text{HO}}$	-473.1	$c_{\text{HO}}$	18896
$a_{\text{OH}}$	-41.37	$b_{\text{OH}}$	-52.64	$c_{\text{OH}}$	1170.9
$a_{\text{OO}}$	-114.77	$b_{\text{OO}}$	3286.78	$c_{\text{OO}}$	1132950
$a_{\text{OO}}$	128.13	$b_{\text{OO}}$	-6597	$c_{\text{OO}}$	1603500
$R_{\text{W}}$	0.098656	$R_{\text{M}}$	0.04938		



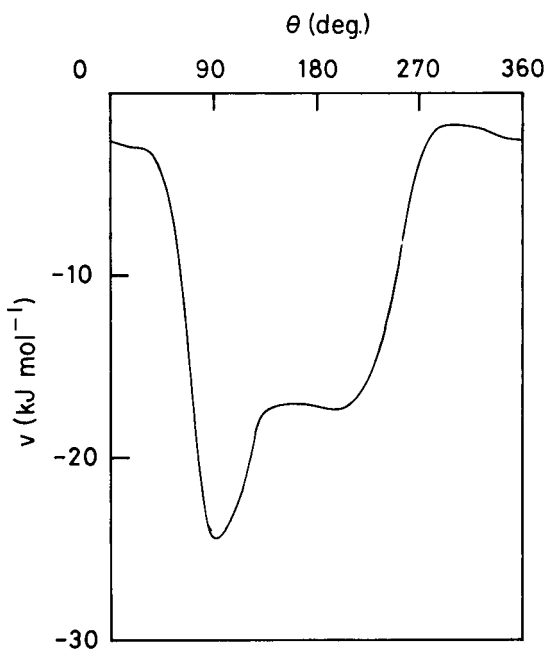


Fig. 2. Angular dependence of the water—methanol pair potential.

values for water in aqueous methanol solution were estimated either as the sum of water—water interaction or the total interaction including that between water and methanol. It was found that the sum of water—water interaction in the hydrophobic region is clearly lower than that in the water—methanol hydrogen bonding region, indicating a hydrophobic hydration of methyl groups. When the water—methanol interaction is taken into account, the potential energy decreases appreciably in the hydrogen-bonded direction, while this is not the case in the hydrophobic direction. Methanol has been classified as a rather hydrophilic solute to water and its dissolution into water has been regarded as substitutional. The present result suggests that the hydrophobic character of the methyl group in methanol is not negligible in dilute aqueous solutions. It is interesting that there is a distinct directionality in the hydration structure around methanol.

The present Monte Carlo data made it possible to evaluate radial distribution functions concerning pair interaction energy and local structure. The detailed results of these analyses will be reported later.

The authors thank the Data Processing Center of Kyoto University for the use of the FACOM M-200 computer and the Computer Center, Institute for Molecular Sciences, for the use of the HITAC M-200H computer and the library program IMS GAUSS70 designed by K. Morokuma et al.

## REFERENCES

- 1 H. Popkie, H. Kistenmacher and E. Clementi, *J. Chem. Phys.*, 59 (1973) 1325.
- 2 W. S. Benedict, N. Geiler and E. K. Plyler, *J. Chem. Phys.*, 24 (1956) 1139.
- 3 R. M. Lees and J. G. Baker, *J. Chem. Phys.*, 48 (1968) 5299.
- 4 G. R. Walsh, *Methods of Optimization*, Wiley, New York, 1975.
- 5 O. Matsuoka, E. Clementi and M. Yoshimine, *J. Chem. Phys.*, 64 (1976) 1351.
- 6 J. C. Owicki and H. A. Scheraga, *J. Am. Chem. Soc.*, 99 (1977) 7413.
- 7 N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth and A. H. Teller, *J. Chem. Phys.*, 21 (1953) 1087.

## Short Communication

---

# TOPOLOGY OF TECTOSILICATE FRAMEWORKS

MITSUO SATO\* and TOSHIHIKO OGURA

*Department of Applied Chemistry, Gunma University, Kiryu, Gunma 376 (Japan)*

(Received 27th July 1981)

**Summary.** The concept of coordination networks is introduced for classifying and deriving possible frameworks of tectosilicates, in which the  $n$ th coordination network is defined as a set of all points from topological distance 0 to  $n$  and all their connection lines. In the second network, it is possible to derive 26 different structures by connection of the points with 3 degrees of freedom; 39 different structures of tectosilicates examined were found to be distributed on the simple coordination networks. The results are compared with those of the secondary building unit criterion proposed by Meier and Breck.

Tectosilicates, in which  $\text{TO}_4$  ( $T = \text{Si}, \text{Al}$ ) are tetrahedrally-linked to form a three-dimensional network, comprise silica, feldspar, feldspathoid and zeolite groups. The framework topology of the zeolites has been investigated by several workers. According to Meier [1] and Breck [2], zeolites can be classified into seven groups on the basis of the secondary building unit (SBU). The SBU criterion can be appreciated because it is a simple and effective geometrical means of understanding the complicated framework structures, but it is unsatisfactory in that various frameworks cannot be derived systematically. Smith [3], Alberti and Gottardi [4, 5], and Sato [6] have developed methods of deriving some of the framework structures systematically. However, many other frameworks have still not been derived. This communication offers an alternative approach for classifying and deriving various kinds of frameworks of tectosilicates.

### *Coordination network*

The tectosilicate framework is a network consisting of points ( $T$  atoms) and lines connecting adjacent points, in which oxygen atoms surrounding  $T$  atoms are usually ignored. The framework has two characteristic features: every point has 4 incident lines, and the framework is extended infinitely in three dimensions. In order to characterize this network, a coordination network around a given point is considered. In a given network, an  $n$ th coordination number around a point can be defined as the total number of points at a topological distance  $n$ . Then an  $n$ th coordination network is defined as a set of all points from topological distances 0 to  $n$  and all their connecting lines. It is obvious that the 0th coordination network comprises one point, and the first coordination network a set of one centering point, 4 adjacent points and 4 connecting lines (Fig. 1, a). In the case of topological

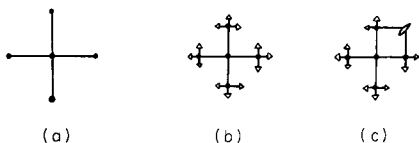


Fig. 1. (a) The first coordination network. (b, c) The second coordination network with (b) 0 connection and (c) 1 connection.

distance 2, the simplest form of the network is tree-like (Fig. 1, b). In this structure, all the 2nd coordination points have 3 degrees of freedom, which means that each T atom coordinates 3 free oxygen atoms around it; these are represented by triangular symbols in Fig. 1. Now, if any two points of these which do not share a common 1st coordination point are connected, one 4-membered ring can be formed (Fig. 1, c) and this is the only possible topological network for one connection. It may be called a 1 connection network. In the figure, one connection point is represented with a symbol  $\mathcal{O}$ , indicating 2 degrees of freedom. From this 1 connection network it is possible to connect two other remaining points with 3 degrees of freedom to form 3 kinds of 2 connection networks. Likewise, 6 kinds for 3 connections, 7 kinds for 4 connections, 5 kinds for 5 connections, and 3 kinds for 6 connections can be obtained. All these are shown in Fig. 2. In addition to the above kind of connection, there are other connection types concerned with points with the same degree of freedom (2), or with different degrees of freedom such as 3 and 2, 3 and 1, and 2 and 1. Some examples are shown in Fig. 3. Now it is convenient to use a topological index to characterize their topological features. Two kinds of topological indices were proposed by Hosoya [7] and Hosoya et al. [8], i.e., the topological index  $Z_G$  and modified topological index  $\tilde{Z}_G$ , which are defined as  $Z_G = \sum_k P(G, k)$  and  $\tilde{Z}_G = \sum_k (-1)^k a_{2k}$ ; here  $P(G, k)$  is a non-adjacent number, i.e., the number of ways of choosing  $k$  independent edges in graph  $G$ , and  $a$  is the coefficient of the characteristic polynomials of  $G$  defined as  $P_G(X) = (-1)^N \det |A - XE| = \sum_k a_k X^{N-k}$  with an adjacency matrix  $A$  and the unit matrix  $E$ .  $Z_G$  values are obtained by Hosoya's method, and  $\tilde{Z}_G$  values by the application of the Frame method to the adjacency matrices. Both topological indices obtained are given in Fig. 2. It is evident that the unmodified topological indices are less sensitive to the structure than the modified ones in this case.

### Classification

All the tectosilicate structures can now be classified on the basis of the 2nd coordination network. The crystal structure consists of the periodic arrangement of unit cells and each point has its own site symmetry, thus coordination networks are obtained for all the points having different site symmetries in a unit cell. The computer program DISTANCE has been developed for that. Thirty-nine different structures examined are shown in Table 1, in which the alphabetical symbols a, b, c, d, e, f, g and h (see Fig. 2) are used instead of topological indices. It is interesting to note that the frame-

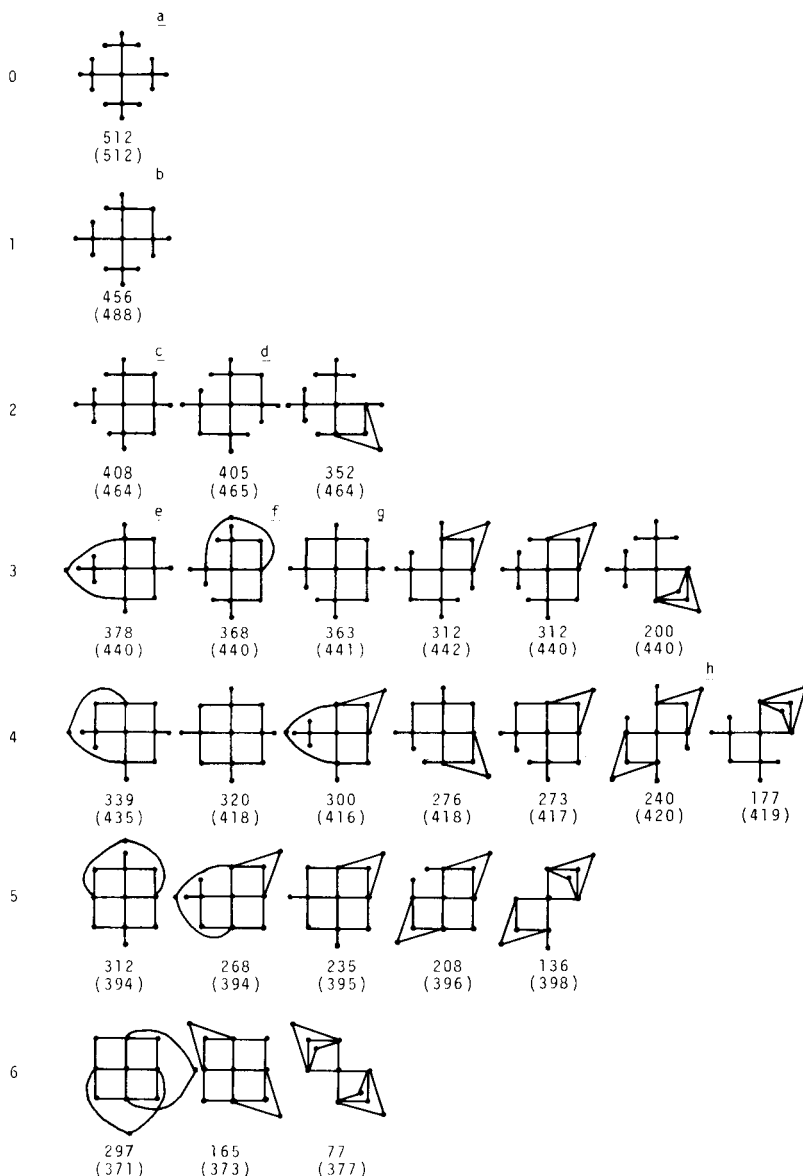


Fig. 2. The second coordination networks formed by connection of points with 3 degrees of freedom. The numbers in the first column are connection numbers. Numerical values added to each network are Hosoya's modified and unmodified (in parentheses) topological indices.

works examined are concentrated on the largest seven modified topological indices, except for natrolite, edingtonite and thomsonite. Group numbers from Breck's classification are listed in the right-hand column. The SBU criterion and the 7 groups are shown in Fig. 4 and Table 2. A comparatively

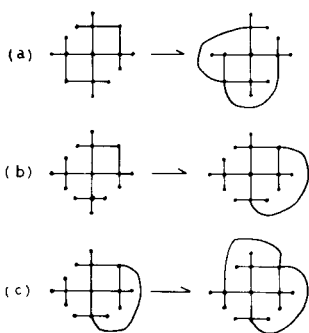


Fig. 3. The second coordination networks formed by connection of two points with different degrees of freedom. Connection with degrees of freedom (a) 2 and 2; (b) 3 and 2; (c) 3 and 1.

good correspondence is obtained between the SBU and the coordination network.

### Discussion

Even at the 2nd distance, there are several kinds of coordination networks. Systematic derivation, as shown in Fig. 2, discloses not only the connective relation between points, but also the relation between 4-membered rings. For instance, there are 3 possible connections for two 4-membered rings, i.e., sharing one corner, one edge and two edges in common respectively, or 5 connection types of three 4-membered rings. The tectosilicate frameworks examined are considered to be classified by these connective relations. Although the SBU criterion does not include any connective relation, the good consistency observed between the SBU and the present coordination networks suggests that they are related. For example, Group 6, complex 5-1 rings, contains only *a* and *b* coordination networks, which suggests that such a 5-membered ring cannot be formed by any combination of 4-membered rings, or it is only possible by some combinations of them which are disconnected from each other. The Group 4, double 6-membered ring, is formed by a *g* coordination network in which three 4-membered rings share one corner in common and one edge in neighboring rings. Because the Group 1, single 4-membered rings, are also formed by the same network, it is questionable to conclude that the *g* network is essential for the formation of such a double 6-membered ring. To clarify the relation between the ring structures and the coordination networks, it is necessary to extend the topological distance  $n = 3$ , i.e., the 3rd coordination network, with which more exact topological distinction of tectosilicate frameworks will be possible.

TABLE 1

The distribution of 39 different tectosilicate frameworks examined on the possible coordination networks<sup>a</sup>

	a	b	c	d	e	f	g	h	R		a	b	c	d	e	f	g	h	R
CR	+									OF			+					+	2
TR	+									OM			+					+	2
NE	+									SO				+					2
QU	+									LA				+					1
FE	+								6	AA				+					1
DA	+	+							6	LE				+			+		2
EP	+	+							6	ZA					+				3
MO	+	+							6	CY						+			
HE	+	+	+						7	HA							+		1
ST	+	+	+		+				7	PA							+		1
SC		+								PH							+		1
BR		+	+						7	GI							+		1
CD		+		+						GM							+		4
YU		+		+					1	CH							+		4
AL			+							FA							+		4
CA			+							ZK							+		4
CO			+							ED								+	5
LO			+	+					2	NA								+	5
ER			+				+		2	TH								+	5
ZL			+				+		4										
AA	Analcime						FA	Faujasite			OF	Offretite							
AL	Albite						FE	Ferrierite			OM	Omega							
BR	Brewsterite						GI	Gismondite			PA	Paulingite							
CA	Cancrinite						GM	Gmelinite			PH	Phillipsite							
CH	Chabazite						HA	Harmotome			QU	Quartz							
CD	Cordierite						HE	Heulandite			SC	Scapolite							
CR	Cristobalite						LA	Laumontite			SO	Sodalite							
CY	Cymrite						ZL	Zeolite L			ST	Stilbite							
CO	Coesite						LE	Levynite			TH	Thomsonite							
DA	Dachiardite						LO	Losod			TR	Tridymite							
ED	Edingtonite						NE	Nepheline			YU	Yugawaralite							
EP	Epistilbite						MO	Mordenite			ZK	ZK-5							
ER	Erionite						NA	Natrolite			ZA	Zeolite A							

<sup>a</sup>Alphabetical symbols a, b, c, d, e, f, g and h correspond to those in Fig. 2, and R means Breck's grouping number of zeolite groups.

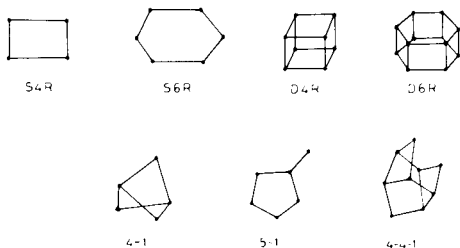


Fig. 4. The secondary building unit networks proposed by Meier [1].

TABLE 2

Breck's classification by the SBU criterion

Group	Secondary building unit (SBU)
1	Single 4-ring, S4R
2	Single 6-ring, S6R
3	Double 4-ring, D4R
4	Double 6-ring, D6R
5	Complex 4-1, $T_5O_{10}$ unit
6	Complex 5-1, $T_8O_{16}$ unit
7	Complex 4-4-1, $T_{10}O_{20}$ unit

## REFERENCES

- 1 W. M. Meier, *Molecular Sieves*, Society of Chemical Industry, London, 1968, p. 10.
- 2 D. W. Breck, *Zeolite Molecular Sieves*, Wiley, New York, 1974, p. 45.
- 3 J. V. Smith, *Am. Mineral.*, 62 (1977) 703; 63 (1978) 960; 64 (1979) 551.
- 4 A. Alberti and G. Gottardi, *N. Jb. Miner. Mh.*, 9 (1975) 396.
- 5 A. Alberti, *Am. Mineral.*, 62 (1977) 1188.
- 6 M. Sato, *Acta Crystallogr., Sect. A*, 35 (1979) 547.
- 7 H. Hosoya, *Bull. Chem. Soc. Jpn.*, 44 (1971) 2332.
- 8 H. Hosoya, K. Hosoi and I. Gutman, *Theor. Chim. Acta*, 38 (1975) 37.



ANALYTICA CHIMICA ACTA, Vol. 133 (1981)  
 (Computer Techniques and Optimization, Vol. 5 No. 4)

AUTHOR INDEX

- Abe, H.  
 —, Yamasaki, T., Fujiwara, I. and Sasaki, S.  
 Computer-aided structure elucidation methods 499
- Abe, H., see Fujiwara, I. 527
- Abe, H., see Miyashita, Y. 603
- Abe, H., see Miyashita, Y. 615
- Angerstein-Kozłowska, H., see Mozota, J. 191
- Azor, M., see Shatkay, A. 183
- Barnett, B., see Mozota, J. 191
- Bartha, I., see Farkas, M. 19
- Broeckaert, I., see Coomans, D. 215
- Broeckaert, I., see Coomans, D. 241
- Bruninx, E.  
 — and van Eenbergen, A.  
 Computerized detection and evaluation of peaks in survey spectra from photoelectron spectroscopy 339
- Burman, J.-O.  
 —, Johansson, B., Morefält, B., Nårfeldt, K.-H. and Olsson, L.  
 Automated inductively-coupled plasma optical emission spectrometry based on a sequential reading monochromator 379
- Cameron, D. G., see Jones, R. N. 555
- Caridi, A. F., see Magallanes, J. F. 203
- Chatt, A., see Tout, R. E. 409
- Chen, J.-H.  
 — and Hwang, L.-P.  
 Reconstruction of mass spectra of components of unknown mixtures based on factor analysis 271
- Clark, R. C., Jr., see Kwan, P. W. 151
- Conway, B. E., see Mozota, J. 191
- Coomans, D.  
 —, Derde, M., Massart, D. L. and Broeckaert, I.  
 Potential methods in pattern recognition. Part 3. Feature selection with ALLOC 241
- Coomans, D.  
 — and Massart, D. L.  
 Potential methods in pattern recognition. Part 2. CLUPOT — an unsupervised pattern recognition technique 225
- Coomans, D.  
 —, Massart, D. L., Broeckaert, I. and Tassin, A.  
 Potential methods in pattern recognition. Part 1. Classification aspects of the supervised method ALLOC 215
- Crandell, C. W., see Smith, D. H. 471
- Csapó, Z., see Szalontai, G. 31
- Dachun, C., see Damo, C. 575
- Daiba, S., see Miyashita, Y. 603
- Damo, C.  
 —, Dachun, C., Teshu, K. and Shaoyu, C.  
 An artificial intelligence system for computer-aided mass spectra interpretation of saturated aliphatic monohydric alcohols 575
- Debska, B.  
 —, Duliban, J., Guzowska-Swider, B. and Hippe, Z.  
 Computer-aided structural analysis of organic compounds by an artificial intelligence system 303
- deMaine, P. A. D.  
 Automatic deductive systems for chemistry 685
- Derde, M., see Coomans, D. 241
- Derendyaev, B. G., see Lebedev, K. S. 517
- de Valk, G. A. J. M., see Jansen, R. T. P. 1
- Domokos, L.  
 — and Frank, I.  
 Orthogonal transformations for feature extraction in chemical pattern recognition 261
- Dorgan, L., see Wilkins, C. L. 637
- Duliban, J., see Debska, B. 303
- Duursma, R. P. J., see Smit, H. C. 283
- Duursma, R. P. J.  
 — and Smit, H. C.  
 User-oriented software for determina-

- tion of the precision of signal-integrating analytical methods 67
- Duursma, R. P. J.
- , Smit, H. C. and Maessen, F. J. M. J.  
Characterization of noise in inductively-coupled plasma emission spectrometry 393
- Eenbergen, A. van, see Bruninx, E. 339
- Esaki, T.  
Development of a graphic program for quantitative drug design 657
- Evans, J. C.  
— and Morgan, P. H.  
Automatic analysis of mixed spectra. Generalised spectral subtraction applied to electron spin resonance spectroscopy 329
- Farkas, M.  
—, Markos, J., Szepesvary, P., Bartha, I., Szalontai, G. and Simon, Z.  
A computer-aided system for organic functional group determinations 19
- Farkas, M., see Szalontai, G. 31
- Frank, I. E.  
—, Pungor, E. and Veress, G. E.  
Statistical decision theory applied to analytical chemistry. Part 1. The statistical decision model and its relation to branches of mathematical statistics 433
- Frank, I. E.  
—, Pungor, E. and Veress, G. E.  
Statistical decision theory applied to analytical chemistry. Part 2. Information and decision in analytical measuring systems 443
- Frank, I., see Domokos, L. 261
- Frazer, J. W., see Herget, C. J. 109
- Foulk, D. S., see Hilliard, L. J. 319
- Fujita, T.  
The *ortho* effect in quantitative structure-activity correlations 667
- Fujiwara, I.  
—, Okuyama, T., Yamasaki, T., Abe, H. and Sasaki, S.  
Computer-aided structure elucidation of organic compounds with the CHEMICS system. Removal of redundant candidates by <sup>13</sup>C-n.m.r. prediction 527
- Fujiwara, I., see Abe, H. 499
- Fujiwara, Y.  
— and Nakayama, T.  
A graph theory data base for storage of chemical structures organized by the block-cutpoint tree technique 647
- Gold, H. S., see Hilliard, L. J. 319
- Goode, S. R., see Matthews, R. J. 169
- Goplen, T. G., see Jones, R. N. 555
- Grabaric, B. S.  
—, O'Halloran, R. J. and Smith, D. E.  
Resolution enhancement of a.c. polarographic peaks by deconvolution using the fast Fourier transform 349
- Gray, N. A. B., see Smith, D. H. 471
- Guzowska-Swider, B., see Debska, B. 303
- Harada, T., see Nishikawa, T. 463
- Herget, C. J.  
— and Frazer, J. W.  
Extension of the Smith predictor 109
- Hijkata, K.  
— and Saito, S.  
Educational equipment in the periphery of a large computer 727
- Hilliard, L. J.  
—, Foulk, D. S., Gold, H. S. and Rechsteiner, C. E.  
Effects of solute-solvent interaction on electronic spectra: A predictive analysis 319
- Hippe, Z.  
Self-adapting computer program system for designing organic syntheses 677
- Hippe, Z., see Debska, B. 303
- Hirose, M., see Hosoya, H. 719
- Hosoya, H.  
— and Hirose, M.  
Graphical representation of the solutions of the Schrödinger equations for a particle in various model potentials 719
- Hwang, L.-P., see Chen, J.-H. 271
- Issahary, D.  
— and Pelly, I.  
A factor analysis study of phosphate beneficiation by calcination 359
- Issahary, D.  
— and Pelly, I.  
A regression analysis study of phosphate beneficiation by calcination 369
- Jacobs, B. E., see Walczak, C. A. 699
- Jansen, R. T. P.  
—, Pijpers, F. W. and de Valk, G. A. J. M.  
Application of pattern recognition for discrimination between routine analyti-

- cal methods used in clinical laboratories 1  
Jinno, K.  
— and Koizumi, T.  
Construction of small data bases for trace element determinations 457  
Jochum, C.  
— and Kowalski, B. R.  
A combined linear and nonlinear factor analysis program package for chemical data evaluation 583  
Johansson, B., see Burman, J.-O. 379  
Johansson, E., see Wold, S. 251  
Jones, R. N.  
—, Cameron, D. G. and Goplen, T. G.  
Computer-aided reduction of absolute infrared intensity measurements by transmission and reflection 555
- Kaminuma, T.  
— and Kurihara, A.  
A data base of data banks for toxicological information 707  
Kanohta, K., see Katagiri, Y. 535  
Katagiri, Y.  
—, Kanohta, K., Yotsui, Y., Nagasawa, K., Okusa, T., Sakai, T. and Tsumura, O.  
Development of a new file search system for nuclear magnetic resonance spectra. Production of an enlarged data base and search test 535  
Kato, Y., see Yamada, A. 421  
Koizumi, T., see Jinno, K. 457  
Komatsui, K., see Moriguchi, I. 625  
Koptuyg, V. A., see Lebedev, K. S. 517  
Kowalski, B. R., see Jochum, C. 583  
Koyama, Y., see Maeda, K. 561  
Krichevsky, M. I.  
Management and querying of morphological, physiological, biochemical and chromatographic data describing microbial strains 747  
Kryger, L.  
Microcomputers in electrochemical trace elemental analysis 591  
Kurihara, A., see Kaminuma, T. 707  
Kwan, P. W.  
— and Clark, R. C., Jr.  
Assessment of oil contamination in the marine environment by pattern recognition analysis of paraffinic hydrocarbon content of mussels 151
- Lebedev, K. S.  
—, Tormyshev, V. M., Derendyaev, B. G. and Koptuyg, V. A.  
A computer search system for chemical structure elucidation based on low-resolution mass spectra 517
- Maas, J. H. van der, see Visser, T. 451  
Maeda, K.  
—, Koyama, Y., Sato, K. and Sasaki, S.  
Combination of analytical spectrometers and spectroscopic data bases 561  
Maessen, F. J. M. J., see Duursma, R. P. J. 393  
Magallanes, J. F.  
— and Caridi, A. F.  
A modified Gran method for determination of equivalence points in potentiometric precipitation titrations 203  
Malinowski, E. R.  
Theory of error applied to pure test vectors in target factor analysis 99  
Malinowski, E. R., see McCue, M. 125  
Markin, R. S., see Wilkins, C. L. 637  
Markos, J., see Farkas, M. 19  
Massart, D. L., see Coomans, D. 215  
Massart, D. L., see Coomans, D. 225  
Massart, D. L., see Coomans, D. 241  
Matherny, M.  
— und Ondáś, J.  
Einsatz von Rechenanlagen in der Emissionsspektrochemie für Aufstellung, Bewertung und Linearisierung der analytischen Eichgeraden 51  
Matherny, M.  
— und Ondáś, J.  
Einsatz von Rechenanlagen in der Emissionsspectrochemie. II. Teil. Aufstellung, Bewertung und Linearisierung der analytischen Eichgeraden 137  
Matthews, R. J.  
—, Goode, S. R. and Morgan, S. L.  
Characterization of an enzymatic determination of arsenic(V) based on response surface methodology 169  
Matsuda, T., see Tanaka, N. 733  
Matsushita, Y., see Moriguchi, I. 625  
McCue, M.  
— and Malinowski, E. R.  
Target factor analysis of infrared spectra of multicomponent mixtures 125  
Moriguchi, I.  
—, Komatsui, K. and Matsushita, Y.  
Pattern recognition for the study of structure—activity relationships. Uses of the adaptive least-squares method and linear discriminant analysis 625

- Miyashita, Y.  
 —, Seki, T., Takahashi, Y., Daiba, S., Tanaka, Y., Yotsui, Y., Abe, H. and Sasaki, S. 603  
 Computer-assisted structure—carcinogenicity studies on polycyclic aromatic hydrocarbons by pattern recognition methods 603
- Miyashita, Y.  
 —, Takahashi, Y., Yotsui, Y., Abe, H. and Sasaki, S.  
 Application of pattern recognition to structure—activity problems. Use of minimal spanning tree 615
- Morefält, B., see Burman, J.-O. 379
- Morgan, P. H., see Evans, J. C. 329
- Morgan, S. L., see Matthews, R. J. 169
- Mozota, J.  
 —, Barnett, B., Tessier, D., Angerstein-Kozłowska, H. and Conway, B. E.  
 The use of a minicomputer for data collection and treatment in the study of electrochemical surface processes 191
- Munk, M. E., see Shelley, C. A. 507
- Nagasawa, K., see Katagiri, Y. 535
- Nakinishi, K., see Okazaki, S. 753
- Nakayama, T., see Fujiwara, Y. 647
- Närfeldt, K.-H., see Burman, J.-O. 379
- Nishikawa, T.  
 —, Ogasawara, I. and Harada, T.  
 Learning-method control applied to microcomputer assisted pH titrations 463
- Nourse, J. G., see Smith, D. H. 471
- Ogasawara, I., see Nishikawa, T. 463
- Ogura, T., see Sato, M. 759
- O'Halloran, R. J., see Grabaric, B. S. 349
- Ohta, K., see Suzuki, M. 209
- Okazaki, S.  
 —, Nakanishi, K. and Touhara, H.  
 Four-M calculations on methanol—water solutions 753
- Okusa, T., see Katagiri, Y. 535
- Okuyama, T., see Fujiwara, I. 527
- Olsson, L., see Burman, J.-O. 379
- Ondáš, J., see Matherny, M. 51
- Ondáš, J., see Matherny, M. 137
- Pelly, I., see Issahary, D. 359
- Pelly, I., see Issahary, D. 369
- Pfeifer, Gy., see Szalontai, G. 31
- Pijpers, F. W., see Jansen, R. T. P. 1
- Plesch, R.  
 Effizienz der Regression in der Röntgenspektrometrie 41
- Pungor, E., see Frank, I. E. 433
- Pungor, E., see Frank, I. E. 443
- Randić, M., see Wilkins, C. L. 637
- Rechsteiner, C. E., see Hilliard, L. J. 319
- Saito, S., see Hijikata, K. 727
- Sakai, T., see Katagiri, Y. 535
- Sakata, M., see Tanaka, N. 733
- Sasaki, S., see Abe H. 499
- Sasaki, S., see Fujiwara, I. 527
- Sasaki, S., see Maeda, K. 561
- Sasaki, S., see Miyashita, Y. 603
- Sasaki, S., see Miyashita, Y. 615
- Sato, M.  
 — and Ogura, T.  
 Topology of tectosilicate frameworks 759
- Sato, K., see Maeda, K. 561
- Schuster, S. M., see Wilkins, C. L. 637
- Seki, T., see Miyashita, Y. 603
- Shaoyu, C., see Damo, C. 575
- Shatkay, A.  
 — and Azor, M.  
 Discrepancies in curve fitting for direct and linearized functions, illustrated with transient electrode potentials 183
- Shelley, C. A.  
 — and Munk, M. E.  
 CASE, a computer model of the structure elucidation process 507
- Siegel, M. M.  
 The use of the modified simplex method for automatic phase correction in Fourier-transform nuclear magnetic resonance spectroscopy 103
- Simon, Z., see Farkas, M. 19
- Simon, Z., see Szalontai, G. 31
- Smit, H. C.  
 —, Duursma, R. P. J. and Steigstra, H.  
 A microprocessor-based instrument for correlation chromatography and data processing 283
- Smit, H. C., see Duursma, R. P. J. 67
- Smit, H. C., see Duursma, R. P. J. 393
- Smith, D. E., see Grabaric, B. S. 349
- Smith, D. H.  
 —, Gray, N. A. B., Nourse, J. G. and Crandell, C. W.  
 The DENDRAL Project: recent advances in computer-assisted structure elucidation 471

- Smith, G. M., see Woodruff, H. B. 545
- Steigstra, H., see Smit, H. C. 283
- Steiner, S., see Wilkins, C. L. 637
- Suzuki, M.
- , Ohta, K. and Yamakita, T.  
Improved sensitivity using a micro-computer for electronic atomic absorption spectrometry with a metal micro-tube 209
- Szalontai, G.
- , Simon, Z., Csapó, Z., Farkas, M. and Pfeifer, Gy.  
Use of IR and  $^{13}\text{C}$ -n.m.r. data in the retrieval of functional groups for computer-aided structure determination 31
- Szalontai, G., see Farkas, M. 19
- Szepesváry, P., see Farkas, M. 19
- Takahashi, Y., see Miyashita, Y. 603
- Takahashi, Y., see Miyashita, Y. 615
- Tanaka, N.
- , Matsuda, T. and Sakata, M.  
Preliminary study of a trace characterization information system 733
- Tanaka, N., see Yamada, A. 421
- Tanaka, Y., see Miyashita, Y. 603
- Tassin, A., see Coomans, D. 215
- Teshu, K., see Damo, C. 575
- Tessier, D., see Mozota, J. 191
- Tormyshev, V. M., see Lebedev, K. S. 517
- Touhara, H., see Okazaki, S. 753
- Tout, R. E.
- and Chatt, A.  
The effect of sample matrix on selection of optimum timing parameters in cyclic neutron activation analysis 409
- Tsumura, O., see Katagiri, Y. 535
- Valk, G. A. J. M. de, see Jansen, R. T. P. 1
- Vandeginste, B. G. M., see Vollenbroek, J. G. 85
- van der Maas, J. H., see Visser, T. 451
- van Eenbergen, A., see Bruninx, E. 339
- Veress, G. E., see Frank, I. E. 433
- Veress, G. E., see Frank, I. E. 443
- Visser, T.
- and van der Maas, J. H.  
Systematic computer-aided interpretation of vibrational spectra 451
- Vollenbroek, J. G.
- and Vandeginste, B. G. M.  
Some considerations on batch arrival and batch analysis in analytical laboratories 85
- Walczak, C. A.
- and Jacobs, B. E.  
A pictorial query language for use with any data base 699
- Wilkins, C. L.
- , Randic, M., Schuster, S. M., Markin, R. S., Steiner, S. and Dorgan, L.  
A graph-theoretic approach to quantitative structure-activity/reactivity studies 637
- Wold, S.
- and Johansson, E.  
Application of SIMCA multivariate data analysis to the classification of gas chromatographic profiles of human brain tissue 251
- Woodruff, H. B.
- and Smith, G. M.  
Generating rules for pairs — a computerized infrared spectral interpreter 545
- Yamada, A.
- , Yoshikuni, T., Kato, Y. and Tanaka, N.  
Computer-aided measurement of kinetic parameters of electrode reactions of cobalt(III)-ammine complexes at mercury electrodes 421
- Yamakita, T., see Suzuki, M. 209
- Yamasaki, A.  
Construction of a data base for cobalt-59 nuclear magnetic resonance spectrometry 741
- Yamasaki, T., see Fujiwara, I. 527
- Yamasaki, T., see Abe, H. 499
- Yoshikuni, T., see Yamada, A. 421
- Yotsui, Y., see Katagiri, Y. 535
- Yotsui, Y., see Miyashita, Y. 603
- Yotsui, Y., see Miyashita, Y. 615

(continued from outside of cover)

The <i>ortho</i> effect in quantitative structure-activity correlations T. Fujita (Kyoto, Japan) . . . . .	667
Self-adapting computer program system for designing organic syntheses Z. Hippe (Rzeszów, Poland) . . . . .	677
Automatic deductive systems for chemistry P. A. D. deMaine (University Park, PA, U.S.A.) . . . . .	685
A pictorial query language for use with any data base C. A. Walczak (Bethesda, MD, U.S.A.) and B. E. Jacobs (College Park, MD, U.S.A.) . . . . .	699
A data base of data banks for toxicological information T. Kaminuma and A. Kurihara (Tokyo, Japan) . . . . .	707
Graphical representation of the solutions of the Schrödinger equations for a particle in various model potentials H. Hosoya and M. Hirose (Tokyo, Japan) . . . . .	719
Educational equipment in the periphery of a large computer K. Hijikata and S. Saito (Tokyo, Japan) . . . . .	727
<i>Short communications</i>	
Preliminary study of a trace characterization information system N. Tanaka, T. Matsuda and M. Sakata (Sendai, Japan) . . . . .	733
Construction of a data base for cobalt-59 nuclear magnetic resonance spectrometry A. Yamasaki (Tokyo, Japan) . . . . .	741
Management and querying of morphological, physiological, biochemical and chromatographic data describing microbial strains M. I. Krichevsky (Bethesda, MD, U.S.A.) . . . . .	747
Four-M calculations on methanol-water solutions S. Okazaki, K. Nakanishi and H. Touhara (Kyoto, Japan) . . . . .	753
Topology of tectosilicate frameworks M. Sato and T. Ogura (Gunma, Japan) . . . . .	759
<i>Author index</i> . . . . .	765

Elsevier Scientific Publishing Company, 1981

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

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

Special regulations for readers in the U.S.A. — This journal has been registered with the Copyright Clearance Center, Inc. Consent is given for copying of articles for personal or internal use, or for the personal or internal use of specific clients, on the condition that the copier pay through the Center the per-copy fee stated in the code on the first page of each article for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. The appropriate fee should be forwarded with a copy of the first page of the article to the Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970, U.S.A. If no code appears in an article, the author has not given broad consent to copy and permission to copy must be obtained directly from the author. All articles published prior to 1980 may be copied for a per-copy fee of US \$2.25, also payable through the Center. This consent does not extend to other kinds of copying, such as for general distribution, resale, advertising and promotion purposes, or for creating new collective works. Special written permission must be obtained from the publisher for such copying. Special regulations for authors in the U.S.A. — Upon acceptance of an article by the journal, the author(s) will be asked to transfer copyright of the article to the publisher. This transfer will ensure the widest possible dissemination of information under the U.S. Copyright Law.

Printed in The Netherlands.

## CONTENTS

<i>Publisher's Note</i> . . . . .	469
The DENDRAL Project: recent advances in computer-assisted structure elucidation D. H. Smith, N. A. B. Gray, J. G. Nourse and C. W. Crandell (Stanford, CA, U.S.A.) . . . . .	471
Computer-aided structure elucidation methods H. Abe, T. Yamasaki, I. Fujiwara and S. Sasaki (Toyohashi, Japan) . . . . .	499
CASE, a computer model of the structure elucidation process C. A. Shelley (Rochester, NY, U.S.A.) and M. E. Munk (Tempe, AZ, U.S.A.) . . . . .	507
A computer search system for chemical structure elucidation based on low-resolution mass spectra K. S. Lebedev, V. M. Tormyshev, B. G. Derendyaev and V. A. Koptuyug (Novosibirsk, U.S.S.R.) . . . . .	517
Computer-aided structure elucidation of organic compounds with the CHEMICS system. Removal of redundant candidates by $^{13}\text{C}$ -n.m.r. prediction I. Fujiwara, T. Okuyama, T. Yamasaki, H. Abe and S. Sasaki (Toyohashi, Japan) . . . . .	527
Development of a new file search system for nuclear magnetic resonance spectra. Production of an enlarged data base and search test Y. Katagiri, K. Kanohta, Y. Yotsui (Tokyo, Japan), K. Nagasawa (Aichi, Japan), T. Okusa (Ichihara-shi, Japan), T. Sakai (Ibaraki, Japan) and O. Tsumura (Chiba, Japan) . . . . .	535
Generating rules for pairs — a computerized infrared spectral interpreter H. B. Woodruff and G. M. Smith (Rahway, NJ, U.S.A.) . . . . .	545
Computer-aided reduction of absolute infrared intensity measurements by transmission and reflection R. N. Jones (Tokyo, Japan), D. G. Cameron and T. G. Goplen (Ottawa, Ont., Canada) . . . . .	555
Combination of analytical spectrometers and spectroscopic data bases K. Maeda, Y. Koyama (Sakuramura, Japan), K. Sato (Tokyo, Japan) and S. Sasaki (Sendai, Japan) . . . . .	561
An artificial intelligence system for computer-aided mass spectra interpretation of saturated aliphatic monohydric alcohols C. Damo, C. Dachum, K. Teshu and C. Shaoyu (Dalian, People's Republic of China) . . . . .	575
A combined linear and nonlinear factor analysis program package for chemical data evaluation C. Jochum and B. R. Kowalski (Seattle, WA, U.S.A.) . . . . .	583
Microcomputers in electrochemical trace elemental analysis L. Kryger (Aarhus, Denmark) . . . . .	591
Computer-assisted structure—carcinogenicity studies on polycyclic aromatic hydrocarbons by pattern recognition methods Y. Miyashita, T. Seki, Y. Takahashi, S. Daiba, Y. Tanaka, Y. Yotsui, H. Abe and S. Sasaki (Toyohashi, Japan) . . . . .	603
Application of pattern recognition to structure—activity problems. Use of minimal spanning tree Y. Miyashita, Y. Takahashi, Y. Yotsui, H. Abe and S. Sasaki (Toyohashi, Japan) . . . . .	615
Pattern recognition for the study of structure—activity relationships. Uses of the adaptive least- squares method and linear discriminant analysis I. Moriguchi, K. Komatsu and Y. Matsushita (Tokyo, Japan) . . . . .	625
A graph-theoretical approach to quantitative structure—activity/reactivity studies C. L. Wilkins, M. Randić, S. M. Schuster, R. S. Markin, S. Steiner and L. Dorgan (Lincoln, NE, U.S.A.) . . . . .	637
A graph theory data base for storage of chemical structures organized by the block-cutpoint tree technique Y. Fujiwara and T. Nakayama (Sakura-mura, Japan) . . . . .	647
Development of a graphic program for quantitative drug design T. Esaki (Nagoya-shi, Japan) . . . . .	657