DISCUSSION

As with many other computerized literature-searching systems, the original manual system could possibly perform as well or better, and at less cost. The decision to implement a computer system was influenced primarily by its novelty and the accessibility of the minicomputer. The system does have speed and growth potential, though, as well as higher accuracy and ease of executing complex search patterns. It is also easily adapted to other calculators such as the Wang 2200 or the IBM 5100.

Developments in computer hardware and software are rapidly followed by advances in the field of computerized information retrieval,[4-7] one of the most recent examples being the CAIN bibliographic search and retrieval service of the U.S. National Agricultural Library;[4] we report for the first time the use of a programmable calculator in this field. The present system can be enhanced immensely by the implementation of the disk drives that are available for this and other minicomputers. A disk would not only speed up the searches but would also enlarge the present capacity of approximately 900 references to at least 150 000.

ACKNOWLEDGMENT

The Department of Chemical Pathology, University of Pretoria where this work was started, is thanked for permission to use their facilities.

REFERENCES AND NOTES

(1) C. P. Bourne, "Methods of Information Handling", Wiley, New York, N.Y., 1966, p 136.
(2) Reference 1, p 38.
(3) C. P. Bourne and D. F. Ford, "A Study of Methods for Systematically Abbreviating English Words and Names", *J. Ass. Comput. Mach.*, **8**, 538–552 (1961).
(4) J. F. Caponio and L. Moran, "CAIN: A Computerized Literature System for the Agricultural Sciences", *J. Chem. Inf. Comput. Sci.*, **15**, 158–161 (1975).
(5) T. L. Isenhour, W. S. Woodward, and S. R. Lowry, "A Rapid Generalized Minicomputer Text Search System Incorporating Algebraic Entry of Boolean Strategies", *J. Chem. Inf. Comput. Sci.*, **15**, 115–118 (1975).
(6) L. S. Rattet and J. H. Goldstein, "Computerizing Scientific Bibliographies", *J. Chem. Educ.*, **45**, 734–736 (1968).
(7) D. U. Wilde and A. C. Starke, "A Chemical Search System for a Small Computer", *J. Chem. Doc.*, **14**, 41–44 (1974).

Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System

H. R. SCHENK and F. WEGMÜLLER*

BASIC, Basel Information Center for Chemistry (Documentation Center of CIBA-GEIGY Ltd., F. Hoffmann-La Roche & Co. Ltd., and SANDOZ Ltd.), CH-4002 Basel, Switzerland

Received January 30, 1976

Using topology, a versatile substructure search has been developed, characterized by total recall and precision and allowing the retrieval of compounds with any given partial structure. A special language which is comprehensible for the chemist is used for the formulation of the query. In conjunction with a specially developed primary screen, the method has been operational since 1972, allowing a routine retrieval of structures in the entire Chemical Abstracts Service Registry II File (2.3 million compounds) at a reasonable cost. The structures retrieved may be viewed as images or presented in the form of nomenclature, and the corresponding abstracts may be obtained by means of a link file.

1. GENERAL SURVEY

Since the late fifties, CIBA Ltd., J.R. GEIGY Ltd. (CIBA-GEIGY Ltd., since 1970), F. Hoffmann-La Roche & Co. Ltd., and SANDOZ Ltd., have been collaborating in various areas of scientific documentation, with particular emphasis on chemistry. Following a joint visit to the Chemical Abstracts Service (CAS) headquarters in Columbus, Ohio, the Basel chemical firms decided on collaboration to develop a computer system for the utilization of CAS data.

The primary purpose of this system was to serve the needs of chemical research and development by conveying information on literature and patents as prepared and made available by abstracting services (i.e., CAS, Derwent, etc.). Data in the form of structures, codes, and text were to be recorded, stored, and utilized in a rational and dependable manner. In addition, internal structural as well as nonstructural input originating within the individual firms had to be processed in a compatible manner by the same system. Special emphasis was placed on the processing of chemical structures. Thus, special consideration had to be given to the heterogeneous nature of the separate data bases, the specific kind of data in question, and the mode of their representation on the storage medium. Search programs had to be adjusted to the peculiarities of the data and had to offer, in addition, the possibility of combined retrievals. Implicit in this concept was the development of easily comprehensible, user-oriented query languages. In the course of realization, special consideration was given to the rational processing of the very large amount of data available (see Figure 6).

To solve these problems, standardized formats like those used by CAS for its Standard Distribution Format (SDF), for example, were applied to the various kinds of data. Incoming files are converted into the corresponding standard format and subsequently transferred to the data base by means of a single, relatively simple program, which is specific for every kind of data. In view of the discrete, modular setup of the entire system, the future incorporation of additional modules to accommodate requirements unknown today should present no problems. Figure 1 gives a survey of the Basel concept, including the preparation of the data base and the retrieval system.

* Correspondence should be addressed to Mr. D. Ligtenberg, BASIC, Secretary, P.O. Box 273, CH-4002 Basel, Switzerland.

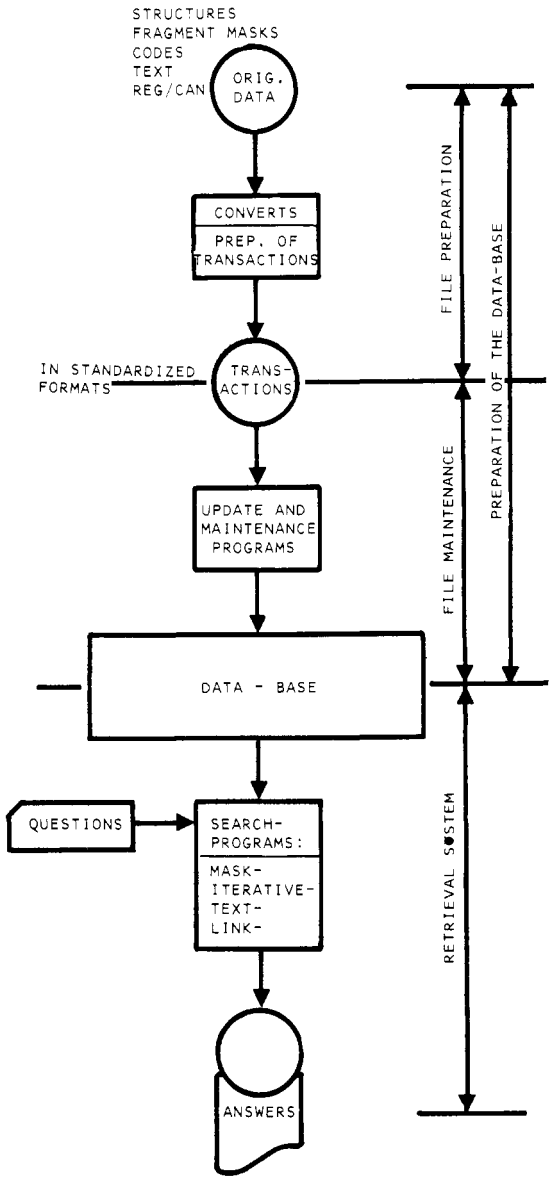


Figure 1. The Basel Information System.

2. CHEMICAL STRUCTURE REGISTRATION AND RETRIEVAL

2.1. The CAS Chemical Registry System. The CAS Chemical Registry System was introduced in 1965. It is based on the process of Gluck[1] for the presentation of chemical structures in the form of topological tables. The numbering according to the Morgan algorithm[2] provides a unique and unambiguous connection table. Connection tables representing structures of compounds being registered are generated by the input of their structural diagrams by means of magnetic tape using a chemical typewriter. The technical progress of graphic displays allows, in addition, the increased use of these displays for structural input. Connection tables obtained in this manner are being matched with those already stored within the Chemical Registry System. Every structure new to the system is assigned a permanent, unique computer-checkable Registry Number which identifies the substance within the CAS data base. Annually, the 14000 periodicals from about 150 countries and patents from 26 countries which are abstracted by CAS contribute about 325000 new compounds to the Registry System.[3] In the beginning of 1975, it comprised 3 million structures.

2.2. The Use of Topology for Substructure Search. Within the areas of research and development, the retrieval of

FRAGMENT NAMES	FRAGMENT MASK DICTIONARY CAS SYMBOL	CAS BITS	BASEL SYMBOL	BASEL BITS
ATOM COUNT	AC	29	AC	29
RING COUNT	RC	13	RC	16
BOND COMPOSITION	BC	125	BC	160
DEGREE OF CONNECTIVITY	DC	24	DC	29
ELEMENT COMPOSITION	EC	182	EC	253
GRAPH MODIFIER	GM	112	GM	
TOTAL A-FRAGMENTS	A	485	A1	487
AUGMENTED ATOMS	AA	817	AA2	1184
TOTAL		1302		1671
SPECIAL SUBSTRUCTURES	SS	93	-	-
LINEAR SEQUENCES	-	-	LS	252
RING BITS	-	-	RB	298
TOTAL FRAGMENT MASKS		1395		2221

Figure 2. Structure of the Fragment Mask Dictionary.

TEST FILE	FRAGMENT TYPES (SEE FIGURE 2)	CANDIDATES PER QUESTION	HITS PER QUESTION	CAND/HIT QUOTIENT	GENERATION COST FACTOR
CBAC 10^4 STRUCT.	A + AA (CAS DICTIONARY)	82	7	11.7	1
	A1 + AA1	63	7	9.0	1
	A1 + AA1 + LS	57	7	8.1	1.8
	A1 + AA1 + SS	44	7	6.3	5.2
	A1 + AA1 + LS + SS	41	7	5.9	6.0
	A1 + AA2	49	7	7.0	1
CAS REGISTRY $5 \cdot 10^4$ STRUCT.	A1 + AA2	324	34	9.5	1
	A1 + AA2 + RB	144	34	4.3	1.5

Figure 3. Fragment Types: efficiency and relative generation costs.

compounds with defined substructures of the Markush formula type is gaining importance. This kind of search is becoming, moreover, increasingly interconnected with questions related to the design of synthetic pathways and to structure/activity correlations. In particular, substructures, which represent defined atom-bond strings and which are comprised entirely or partly of one ring, several rings, or a chain in addition to a ring, are frequently impossible to retrieve by means of fragment codes or linear notations. The "total recall and precision" requirement, indispensable for searching in the steadily growing file of known chemical structures, is satisfied by the use of topology. An iterative atom-by-atom and bond-by-bond match in the connection tables, however, is very uneconomical. To lower the cost of substructure searching in very large data bases like those of the CAS Chemical Registry System (cf. Figure 6), a preliminary selection is necessary. Various screening procedures are already known.[4-12]

The CAS fragment mask procedure, which uses a CAS-developed algorithm to generate a multiplicity of fragments from the connection tables, satisfied our needs. In a further step, a comparison with a dictionary containing a limited number of chemically significant fragments, yields the corresponding fragment masks in the form of a bit string.

The list of CAS-defined A and AA fragments[13] has been adjusted to our requirements and expanded (→ A1 respectively AA1 respectively AA2) (see Figure 2). We have, furthermore, defined 252 bits of Linear Sequences (this type of fragment has been provided for by CAS, but not contained in the corresponding dictionary) and have newly created 300 ring bits which take into account the size, kind, degree of saturation, fusion, and substitution of all rings contained in the registered structures.

We have tested the screening effect of various groups of fragments on two test files of about 10000 and 50000 structures each,[14] using 100 actual questions. The relation,

number of candidates / number of hits[15], has been used as a measure of efficiency. The original CAS dictionary (A + AA) yielded a number of candidates / number of hits ratios of 11.7. The factor 1.0 (for the generation cost of A + AA fragments) has been used as a basis for the comparison of the corresponding cost for different groups of fragments (Figure 3). The utilization of Basel A1 fragments in conjunction with a first extension of the AA fragments (=AA1), together with CAS Special Substructures (SS) and the Linear Sequences (LS) defined by us, has reduced the above ratio by about one-half. On the other hand, the generation cost has increased sixfold. Since this increase was primarily due to the generation of SS and LS fragments, we preferred to refrain from using them, trying to achieve an acceptable candidates / hits ratio at a possibly low generation cost through a further expansion of AA1 (AA2) fragments. The best screening effect (number of candidates / number of hits ratio of 4.3) is achieved by the addition of ring bits to the A1 + AA2 dictionary. Since, in this case, the generation cost is 50% higher, we are refraining from the use of ring bits for the time being.

2.3. The Basel Approach to the Substructure Search. Substructure search as implemented in Basel comprises a mask search as a primary screen, followed by an iterative search, and completed by a link between the retrieved structures (Registry Numbers) and the CAS data base. For this entire sequence, three files are required: a Fragment Mask File for the primary selection, the Chemical Registry File containing the structures in the form of their connection tables, and a Reference File linking these with the abstract numbers. In contrast to the CAS practice of using the Chemical Registry System for the purpose of registration of compounds and preparation of indexes, our primary goal is to use it for substructure retrievals. This implies a special processing of the files which we receive from CAS.

In the spring of 1971, CAS put at our disposal its 1970 Registry File, containing, at that time, about 1.5 million structures in two versions. The first of these versions, which was in nested form, was used by CAS for registration purposes and was sorted according to the connection tables. The other version was in unnested form, i.e., in Registry Number sequence.[16] (Nesting is a file compaction technique whereby certain information that is identical in adjacent records is not repeated after its first citation.) For economic and technical reasons, the unnested version of the file (sorted according to Registry Numbers and allowing the same sequence of data in the Registry File and in the corresponding Mask File) is being used by us for substructure searching. In addition, the subsequent linking of the Registry Numbers with the CAN's (Chemical Abstract Numbers) is greatly facilitated.

For CAS, as well as for us, the most mutually convenient procedure for the updating of the Basel structure file is to match, in an initial step, the newly delivered, complete CAS Registry File with the previous version (Figure 4). In this step, the (unnested) new additions are extracted. At the same time, the Registry Numbers of amended structures, that is, those present in the old file but absent in the new one, are written out separately and marked accordingly. All these transactions are subsequently sorted according to Registry Numbers and constitute the transaction file for updating. The Fragment Mask File based on the Basel dictionary was prepared for the first time using the CAS Structure File from the first quarter of 1972 and contained at that time 1.9 million structures. The generation was performed using partly modified CAS programs. We succeeded in greatly compressing the Fragment Mask File using various measures such as rearranging the bits in the order of statistical frequency (cf. Figures 11 and 12), truncating the records following the last bit present, and increasing the blocking factor. In the course

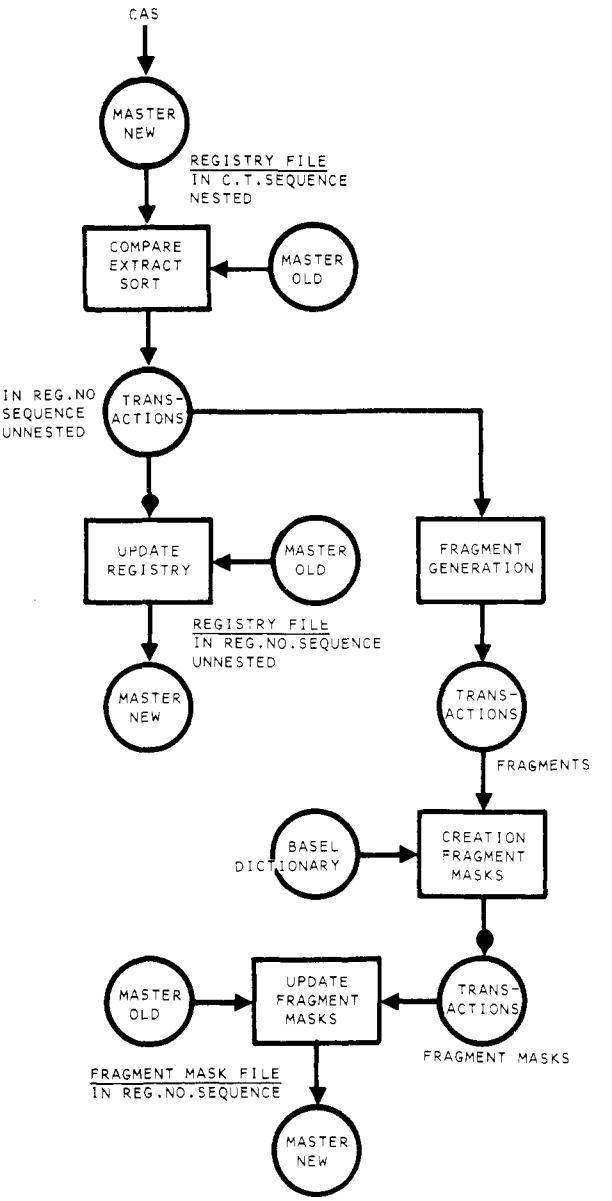


Figure 4. Structural data: conversion and updating.

of updating the Fragment Mask File dictionary, fragments and fragment masks are generated from all newly added structures and are used to update the master file. Simultaneously, fragment masks eliminated by matching the old with the new Registry File are removed from the master file (Figure 4).

As output, the substructure search yields lists of Registry Numbers which allow an unequivocal identification of the structures in question, but which give no direct indication about their occurrence in the literature or about their physical, chemical, or biological properties in general. A link file is required to connect these Registry Numbers with the CAS abstract numbers. To create this REG/CAN (Registry Number to Chemical Abstract Number) File, CAS gave us an intermediate file, which was used in its editorial processing and which contains REG/CAN data for the period 1967 to 1973. For the same data of the preceding 1965/66 period, we obtained a specially prepared Archival Link File.

For updates in the Chemical Registry System, CAS distinguishes between two kinds of transactions: in one case, a structure record is deleted and the registration noted as no longer valid when a review has shown that the registered substance is not a unique chemical entity; in the other case, a structure record is cross-referred to another record in the file when it is recognized that two different structure records

NO. OF REG.NOS	NO. OF CAN PER REG.NO.
1'788'506	1
318'091	2
95'881	3
41'221	4
21'842	5
13'086	6
8'849	7
25'422	8 - 15
10'737	16 - 31
8'379	32 - 127
3'173	128 - 1'023
438	1'024 - 16'383
11	BEYOND 16'383

Figure 5. REG/CAN statistics.

FILE / NO. OF	RECORDS	TAPES (1600 BPI)
REGISTRY STRUCTURE	2'340'231	10
FRAGMENT (1 MASTER + 1 UPDATE)	2'533'294	53
FRAGMENT MASK (TRUNCATED)	2'340'231	11
(BIT SORTED)	2'340'231	9
REG/CAN	6'604'994	3

Figure 6. Size of the Basel data base.

QUESTION:

THE N-HETEROCYCLE IS ISOLATED (NO FUSION ALLOWED), IS AT LEAST 5-MEMBERED AND CONTAINS AT LEAST 1 DOUBLE BOND AND, AT THE MOST, A SECOND HETEROATOM (N, O OR S). THE RING MAY BE SUBSTITUTED. THE BENZENE RING IS ISOLATED (NO FUSION ALLOWED) AND MAY CARRY AT MOST 2 ADDITIONAL SUBSTITUENTS.

FORMULATION:

```
CYCLUS-A        HETEROCYCLE WITHOUT BRIDGES
                SIZE MIN 5
                ATOMS PERMITTED C, N, O, S
                ATOMS REQUIRED MAX 2 HET
                BONDS REQUIRED MIN 1 #2
                REL Ø
                PART N<1>

CYCLUS-B        BENZENE
                REL Ø
                CHEL MAX 3

ALKYL           C-CHAIN-AREA
                SIZE 1 - 8
                BONDS PERMITTED :1
                CHEL 1

STRUCTURE       CYCLUS-A<1> :1 CO1 :1 N<2> :1 CYCLUS-B;
                <2> :1 ALKYL
```

EXPLANATION:

```
REL             RING ELEMENTS
CHEL            CHAIN ELEMENTS
#1              RING SINGLE BOND
#2              RING DOUBLE BOND
:1              CHAIN SINGLE BOND
$               ALL KINDS OF BONDS
```

Figure 7. Query language: question and formulation.

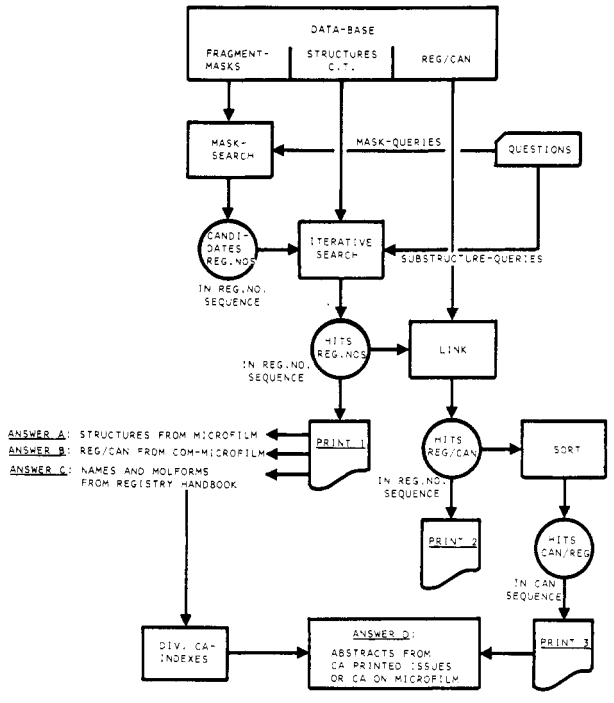


Figure 8. The retrieval system.

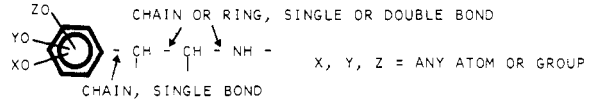


Figure 9. Example of a substructure search: question.

BIT NO.	DESCRIPTION OF 'AUGMENTED ATOM' TYPE	FREQUENCY
98	ONCE: C C N (ALL KINDS OF BONDS ARE ALLOWED BETWEEN THESE ATOMS)	59.73%
105	6 TIMES: C *4 C *4 C (WITH EQUALIZED AROMATIC BONDS BETWEEN THESE C-ATOMS)	57.62%
163	ONCE: C *4 C *4 C -: C (THREE C-ATOMS WITH EQUALIZED AROMATIC BONDS BETWEEN THEM, THE CENTRAL OF THESE ATOMS ATTACHED TO A FOURTH C-ATOM BY A SINGLE CHAIN BOND)	36.78%
696	3 TIMES: C *4 C *4 C -1 O (THREE C-ATOMS WITH EQUALIZED AROMATIC BONDS BETWEEN THEM, THE CENTRAL OF THESE ATOMS ATTACHED TO AN O-ATOM BY A SINGLE CHAIN BOND)	2.30%

Figure 10. Example of a substructure search: Fragments used. A central "augmented" atom is specified first, followed by the designation of bonds and corresponding atoms which are attached to it. "Any kind" of bond is denoted by a blank, a chain bond by a hyphen, and a cyclic bond by an asterisk. Within the chain and cyclic bonds, single, double, and triple bonds are designated by a 1, 2, and 3 and equalized or delocalized bonds by a 4, respectively. Completely conjugated cyclic (for example, aromatic) bonds are also designated by a 4 (cf. augmented atoms = fragment type AA on the dictionary page in Figure 11).

have been unintentionally established for the same substance. Multiple registrations in the CAS file typically result from different structural representations in the original scientific literature for the same substance. All these changes are recorded in the special Deleted Registry Number File. Since these Registry Numbers will no longer be included in future CAS products, they must be given special consideration in the course of processing for our REG/CAN File, from which deleted Registry Numbers are not removed but are flagged. This allows the retrieval of entries made before the update. Occasionally, a structure record is cross-referred more than once. In such cases the related Registry Numbers are matched, the last valid one is determined, and all CAN's are assigned to the presently valid Registry Number. Since these changes disturb the numeric sequence, the REG/CAN File must be sorted again according to Registry Numbers. The REG/CAN statistics (Figure 5) make evident that 76% of all compounds are referenced to one abstract only and that 90% of all compounds are referenced to no more than two abstracts.

Figure 6 presents information about the size of our data base with the three files required for a complete substructure search.

2.4. Description of a Substructure Search. To use the iterative search programs which were put at our disposal by CAS, the substructure must be formulated atom by atom with

```
DCA-SYSTEM        FRAGMENT SEARCH DICTIONARY        PAGE 019
DATE: 30.10.75

BIT-N  TYPE BIT-O  CNT   SCREEN ITEM
0656   AA   1366   002   N -1C -2O -2O
0657   AA   1302   001   N * C * C * C
0658   AA   0902   001   C -1C -3N
0659   AA   1026   001   C *1C *1S
0660   AA   1373   003   N   N
0661   AA   0941   002   C -1C -1N -2O                $
0661   AA   0941   002   C -1C -2N -1O               $
0662   AA   1586   002   S -1C
0663   AA   0829   001   C * C * C * S
0664   AA   1127   001   C   N   N   S
0665   AA   0779   002   C *4C *4C *1N
0666   BC   0189   030   -
0667   AA   0512   001   C - BR* C
0668   AA   0910   001   C * C * N - N
0669   AA   1422   001   N - S
0670   AA   1467   003   O   P
0671   AA   1154   001   C - N - O - O
0672   AA   0763   001   C *1C *1C *1N
0673   AA   1423   001   N -1S
0674   AA   0787   001   C *4C -1C *4N
0675   AA   1009   001   C   C   P
0676   AA   0942   001   C   C   N   S
0677   AA   1639   001   S   N   O   O
0678   AA   1348   001   N -2C -1N
0679   AA   1035   001   C -1C *1S
0680   AA   1029   001   C *2C *1S
0681   DC   0063   018   3
0682   AA   0522   001   C - BR* C * C
0683   AA   1194   001   C * O * O
0684   AA   1436   009   O   C
0685   AA   0915   001   C - C * N * N
0686   AA   0602   004   C -2C
0687   AA   0757   005   C   C   C   N
0688   AA   1553   001   P   O   O   O
0689   EC   0281   040   C
0690   AA   1237   001   C -2S
0691   AA   0781   001   C *4C *4C *4N
0692   AA   1169   001   C - N - S
0693   AA   0914   001   C *1C *2N -1N               $
0693   AA   0914   001   C *2C *1N -1N               $
0693   AA   0914   001   C *4C *4N -1N               $
0694   AA   1610   001   S - C - N
0695   AA   0732   002   C *4C *4C -1CL
0696   AA   0818   003   C *4C *4C -1O
0697   AA   1626   001   S   C   O   O   O
0698   AA   0892   003   C *4C *4N
0699   AA   1611   001   S   C   N   O   O
0700   EC   0447   000   CS                          $
```

a

OLD BIT #	NEW BIT #	FREQUENZ	%
800	423	143502	7.41
801	701	44186	2.28
802	615	60248	3.11
803	373	187765	9.69
804	568	75458	3.89
805	777	30983	1.60
806	958	13443	.69
807	428	140450	7.25
808	729	39274	2.03
809	1012	10920	.56
810	831	25117	1.30
811	1199	4479	.23
812	518	98020	5.06
813	739	36857	1.90
814	550	83048	4.29
815	738	37730	1.95
816	267	343394	17.72
817	448	129900	6.70
818	696	44659	2.30
819	822	25903	1.34
820	313	256728	13.25
821	419	144867	7.48
822	492	113374	5.85
823	1011	10935	.56
824	999	11824	.61
825	1407	1281	.07
826	779	30773	1.59
827	1213	4262	.22
828	356	208638	10.77
829	663	51345	2.65
830	482	116306	6.00
831	731	39171	2.02
832	1503	532	.03
833	1181	4900	.25
834	1136	5830	.30
835	1484	602	.03
836	1529	404	.02
837	1650	20	.00
838	1154	5465	.28
839	982	12593	.65

b

Figure 11. Fragment mask dictionary: (a) coding section, (b) bit statistic.

indication of the allowed or prohibited bonds. Since this kind of formulation has proved to be very time consuming and cumbersome, we have developed for our own search program a special query language which eliminates these difficulties, while at the same time is comprehensible by the chemist (Figure 7). The retrieval program based on this query language has been operational since 1972.

For a substructure search in the entire Registry File, a mask search is performed as a primary screen. For a formulation of a query, only those bits which are selective enough, that is, those with a low statistical frequency, are used (cf. Figure 12). The mask search yields the Registry Numbers of the candidates. In the following iterative search, only the connection tables of the candidates are examined (Figure 8).

In comparison to a purely iterative search in a file of 2.3 million compounds, the screens containing our dictionary, in combination with a specially developed optimized tape access, can reduce the computer costs by 80–95%. The savings depend on the specificity of the dictionary with respect to the query, i.e., on the number of candidates retrieved. The following example is intended to illustrate an entire cycle of a retrieval, including all possibilities of evaluation.

Searched: All compounds containing an isolated benzene ring (no fusion allowed), with at least four substituents and represented by the Markush formula in Figure 9.

```
CYCLUS-A    BENZENE
            REL Ø
            CHEL MIN 4
STRUCTURE   CYCLUS-A :1 O;
            CYCLUS-A :1 O;
            CYCLUS-A :1 O;
            CYCLUS-A :1 CH1 $ CH1 $ NH1
```

Figure 12. Example of a substructure search: query language formulation. The requirement of at least four substituents at the ring would not itself be necessary, since these are being specified under "STRUCTURE". It does shorten the search time considerably, however, since all structures with less than four substituents are rejected and thus withheld from any further examination.

For the mask search screen (cf. Figure 8) the Fragment Mask bits in Figure 11 are applied, using AND logic (Figure 10). Bit No. 696, which is the most selective according to its frequency, would alone produce somewhat more than 42000 candidates. The utilization of all four bits mentioned gave 17352 candidates. The CPU time of the subsequent iterative search was close to 20 s. In view of this, it is not absolutely necessary to decrease the number of candidates further by using additional Fragment Mask bits.

For the iterative search, the query is formulated according to Figure 12 (abbreviations used are explained in Figure 7). The substructure search results in a hit list containing Registry Numbers of the 483 relevant structures (Figures 8 and 13,

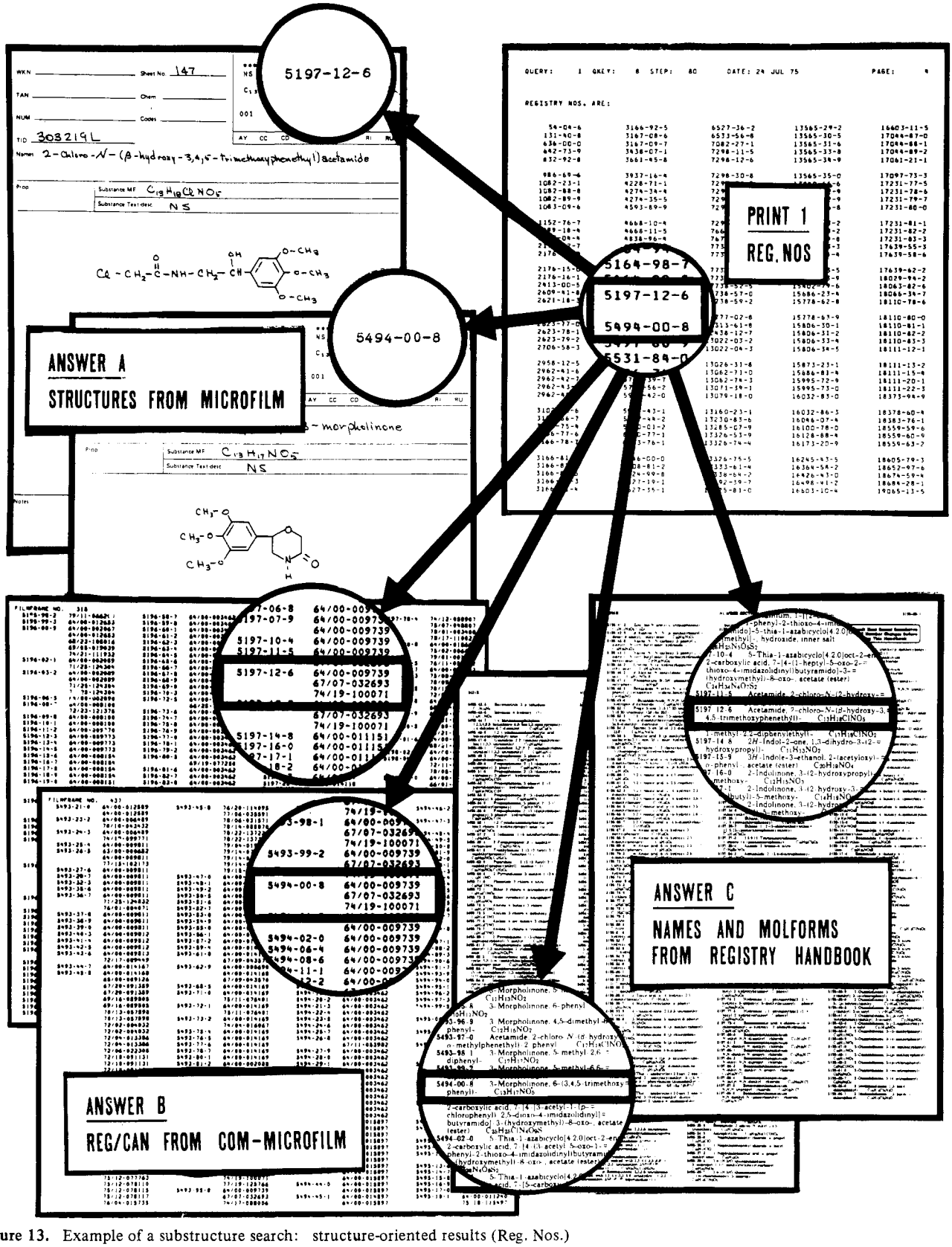


Figure 13. Example of a substructure search: structure-oriented results (Reg. Nos.)

Print 1). If the aminoethyl or aminoethenyl side chain is allowed to be part of a ring in fusion with the benzene ring, the search yields 610 compounds including, in addition, indoles, isoquinolines and further fused systems, and the corresponding hydrogenated derivatives.

Using Print 1, the structural formulas of the retrieved compounds may be viewed or copied from the Registry microfilm (Figure 13, Answer A), which constitutes a satisfactory result from the structural point of view. In many cases, however, a reference to the chemical, physical, or biological properties of these compounds is required, which means a further effort in order to retrieve the pertinent secondary or primary literature.

Although by using Print 1 in conjunction with the Registry Handbook (Figure 13, Answer C) a manual search of this kind in the various CA Indexes is possible, an effort-saving,

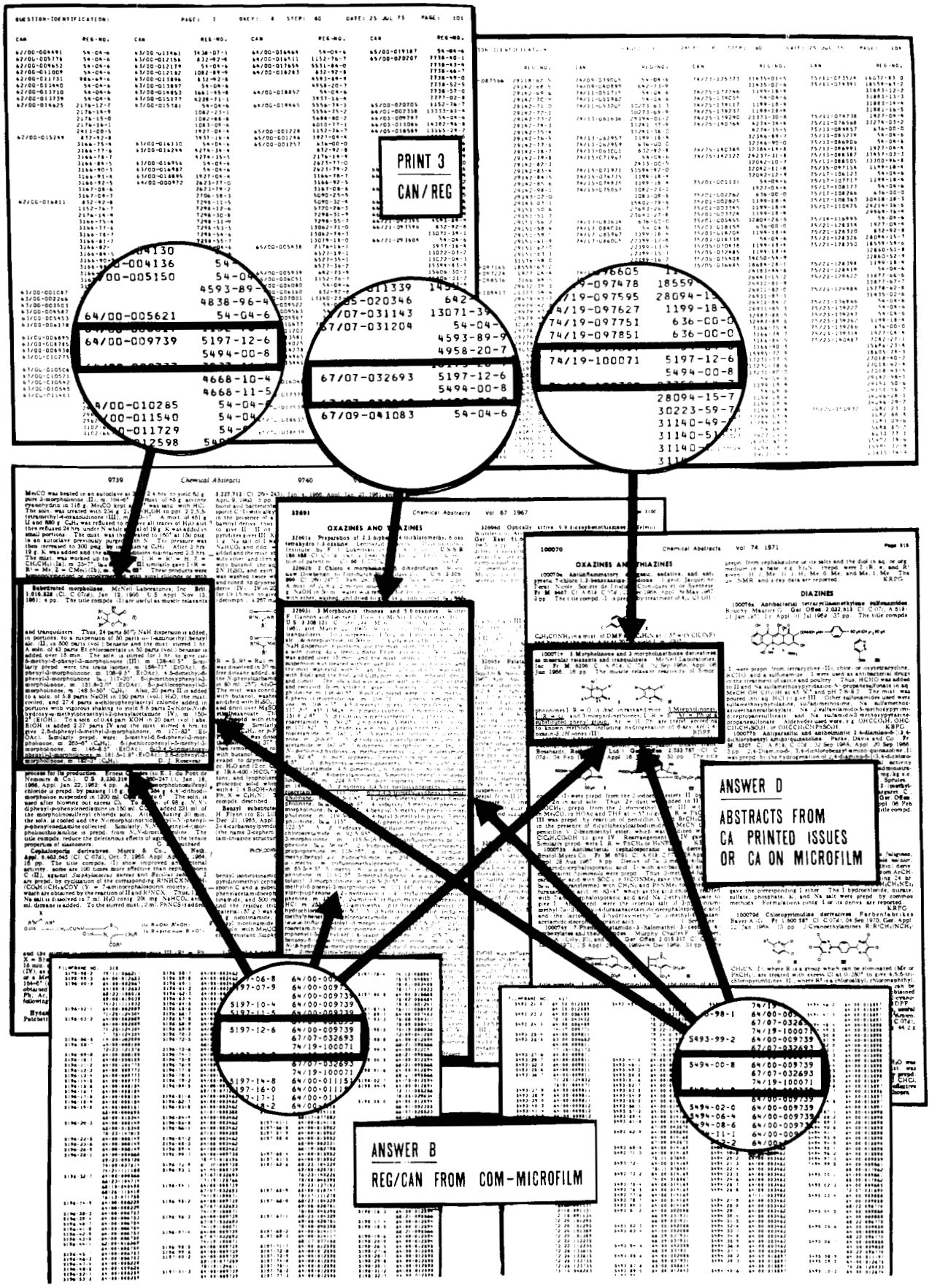
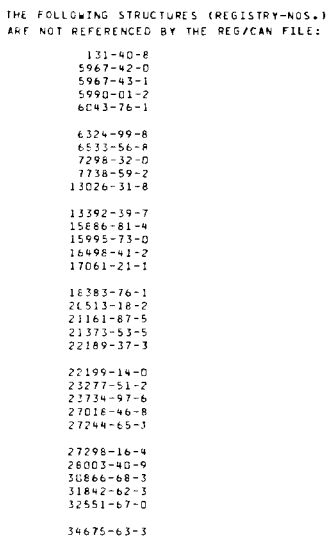


Figure 14. Example of a substructure search: abstract-oriented results (CAN).

computer-based service is offered. The link file mentioned before may be used in two ways: the CAN's belonging to the Registry Numbers of Print 1 may be retrieved from the REG/CAN microfilm[17] (Figure 13, Answers B), or a corresponding list (Figure 8, Print 2) can be printed by the computer via the link search program.


THE FOLLOWING STRUCTURES (REGISTRY-NOS.)
ARE NOT REFERENCED BY THE REG/CAN FILE:

131-40-8
5967-42-0
5967-43-1
5990-01-2
6043-76-1

6324-99-8
6533-56-8
7298-32-0
7738-59-2
13026-31-8

13392-39-7
15686-81-4
15995-73-0
16498-41-2
17061-21-1

18383-76-1
20513-18-2
21161-87-5
21373-53-5
22189-37-3

22199-14-0
23277-51-2
23734-97-6
27018-46-8
27244-65-3

27298-16-4
28003-40-9
30866-68-3
31842-62-3
32551-67-0

34675-63-3

Figure 15. Example of a substructure search: structure results without REG/CAN reference.

In actual practice, a rearrangement of Print 2 into the CAN sequence (all Registry Numbers pertaining to a single CAN being listed together, as in Figure 14, Print 3) has proved in many cases to considerably simplify further evaluation, since abstracts may be viewed in numerical sequence, which eliminates the redundant reading of the same abstracts in connection with different Registry Numbers.

The Registry File also covers compounds from such sources which are not being referenced in CA (i.e., the Merck Index, Beilstein, etc.). Because of this, preceding Print 2 or 3 and Answer B, a list of those retrieved Registry Numbers which pertain to compounds of this kind is printed (Figure 15).

3. FURTHER ASPECTS

The methods described above make a correlation of structure with the general text of the abstracts possible; a detailed definition and evaluation of nonstructural concepts, however, calls for an additional intellectual and manual effort. The biological context, in particular, is of eminent importance for the pharmaceutical industry. Because of this, in 1973 (before the advent of the REG/CAN File) we endeavored to realize a link between the structural and those biological concepts which are expressed in the text. Since the Chemical-Biological Activities (CBAC) file contained Registry Numbers in the abstract text from the very beginning (1965), it was possible to extract them with the corresponding abstract numbers to create a special REG/CAN file for this service. From the beginning of 1973, we were able to link the results of a substructure search with biological data by means of a subsequent text search. A retrieval of this kind yields a highly specific answer in the form of a printout containing the relevant CBAC abstracts.

An analogous service may, in principle, also be arranged for the special documentations which have been introduced by CAS since the beginning of 1975 (Energy, Materials, Food and Agricultural Chemistry, Ecology and Environment, and with some limitations also Polymer Science and Technology[18]).

3.2. Internal Compounds. As mentioned in the introduction, our own internal compounds are also being registered using the CAS process. Chemical typewriters of the DURA 1041, DATICA, and INVAC MDS types are used for the input. The advantage of doing so is that structures published in the literature as well as our internal compounds may be searched using the same query and the same programs. At the same time, the retrieved structural data of our internal compounds may be linked with the corresponding biological test results.

3.3. Chemical Registry System III. The initial version of the CAS Chemical Registry System, referred to as Registry I, was established in 1965. This experimental registration of fully defined organic chemical substances proved the viability and validity of the registration concept. In 1968 the scope of the system was increased as additional classes of substances were handled, and this system, referred to as Registry II, began to be integrated into the CAS indexing operation.

In the beginning of 1974, CAS changed from the Registry II to the Registry III System which is characterized, among other things, by a separate treatment of chains and ring systems and by new rules covering tautomerism. The problem of making Registry II and Registry III compatible is the subject of close collaboration between CAS and Basel specialists. The results will be published at a later date.

4. CONCLUSION

The three Basel chemical firms have developed in common an integrated information system, which, owing to its use of topology, opens new possibilities of questioning for substructure and structure search among all published structures registered since 1965 by CAS and among those structures of registered internal compounds. Characterized by total recall and precision, this kind of search may also be coupled with the retrieval of nonstructural data, thus offering the scientist a new and efficient aid.

ACKNOWLEDGMENT

The Basel group thanks CAS for placing at its disposal all technical records and computer programs for conversion of structural input into connection tables, generation of fragments, preparation of the Fragment Mask File, execution of the iterative search, and for the utilities programs as well as for the current delivery of the Registry, REG/CAN, and Archival Link Files. We are also greatly indebted to CAS for the most valuable stimulus provided by numerous personal contacts, technical advice, and professional discussions and should like to acknowledge specially the most kind and generous consideration shown for our own, specific industrial needs. The subject of this publication has been contributed to by the Scientific Documentation departments of CIBA-GEIGY Ltd. (A. Jacob, B. Schmidt), F. Hoffmann-La Roche & Co. Ltd. (G. Ernst, F. Wegmüller), and SANDOZ Ltd. (H.K. Kaindl, H.R. Schenk) and with the cooperation of J. Fillinger, A. Gehring, U. Hegi, and others from the Data Processing Departments of the three companies. This team is grateful to the managements of the individual firms for the goodwill and the resources which made this work possible.

REFERENCES AND NOTES

(1) J. D. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont", *J. Chem. Doc.*, **5**, 43 (1965).
(2) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107 (1965).
(3) (a) "CAS TODAY, Facts and Figures About Chemical Abstracts Service", Chemical Abstracts Service, Columbus, Ohio, 1974, 32 pp; (b) P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.*, **16** (2), 111 (1976).
(4) G. W. Adamson et al., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. V. More Detailed Cyclic Fragments", *J. Chem. Soc., Perkin Trans. 1*, in press; "IV. Cyclic Fragments", *ibid.*, 863 (1973); "III. Statistical Association of Fragment Incidence", *ibid.*, 2428 (1972); "II. Atom-Centered Fragments", *J. Chem. Soc. C*, 3702 (1971); "I. Non-Cyclic Fragments", *ibid.*, 990 (1970).
(5) G. W. Adamson et al., "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base", *J. Chem. Doc.*, **13**, 159 (1973).
(6) G. W. Adamson et al., "An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments", *J. Chem. Doc.*, **14**, 44 (1974).

(7) G. W. Adamson et al., "Relationship between Query and Data-Base Microstructure in General Substructure Search Systems", *J. Chem. Doc.*, **13**, 133 (1973).
(8) G. W. Adamson et al., "Distributions of Fragment Representations in a Chemical Substructure Search Screening System", *J. Chem. Doc.*, **14**, 72-74 (1974).
(9) V. H. R. Bragg et al., "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems", *J. Chem. Doc.*, **10**, 125 (1970).
(10) A. J. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147-152 (1975).
(11) M. F. Lynch in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke et al., Ed., Wiley, New York, N.Y., 1974, pp 31-53.
(12) E. Meyer, "Topological Search for Classes of Compounds in Large Files—even of Markush Formulas—at Reasonable Machine Cost", in ref 11, pp 105-122.
(13) AA = Augmented Atoms (see caption to Figure 10).
(14) Every 4th of the 40 000 CBAC structures and every 36th connection table contained in the 1972 Registry file (1.9 million structures at that time).
(15) Candidates = structures resulting from the mask search. Hits = structures resulting from the iterative search.
(16) "System Documentation for the Chemical Abstracts Service Registry System", Chemical Abstracts Service, Columbus, Ohio, 1968.
(17) Obtained from the REG/CAN file by Computer Output on Microfilm (COM).
(18) Polymer structures recorded in Registry II cannot be handled by our iterative search.

An Empirical Method of Structure–Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation

GEORGE W. ADAMSON* and DAVID BAWDEN

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received February 17, 1976

A method of substructural analysis for structure–property correlation and property prediction allowing representation of the effects of positional isomerism and substituent interaction is described. Rate constants for the bromination of 44 substituted benzenes are correlated by means of multiple regression analysis using sets of structural features derived automatically from Wiswesser Line Notation. The best set of structural features gives a multiple correlation coefficient >0.999. Property predictions are simulated for 24 compounds, with up to 5 substituents. The technique could be carried out automatically with large machine-readable structure–property files, and may be generally applicable to the properties of substituted cyclic compounds.

The correlation of properties with chemical structure and the consequent prediction of unknown values has long been a major goal of physical organic chemistry and is currently of considerable importance in such fields as drug design.[1] Several approaches to this problem have been employed, varying from purely empirical to highly theoretical.

Quantum mechanical methods have been applied to practical problems[2] but have not yet been applied very widely to the correlation of structure and biological activity. The recently developed MINDO/3 methodology[3] has been suggested to be of wider applicability, while calculations based on molecular mechanics[4] have been used successfully in some cases.

Semiempirical correlation methods, generally known as linear free energy relationships, have been very widely employed. The well-known Hammett equation and its derivatives[5,6] have been principally applied to structure–reactivity correlation for organic compounds. Its main successes have been in summarizing and clarifying experimental data and in aiding the elucidation of reaction mechanisms, though some property predictions have been made using such methods.[7] The Hansch methodology,[8] which aims to correlate biological activities with physicochemical molecular properties, has also been widely used.

Empirical relationships have long been used for the estimation of unknown property values, particularly for thermodynamic quantities.[9] More recently statistical modelling has been employed for the prediction of biological properties,[10] while pattern recognition techniques have qualitatively predicted Hammett values.[11]

A range of methods generally described as "substructural analysis", involving the correlation by computerized statistical analysis of structural features with property values, have enabled both qualitative and quantitative predictions of various biological properties.[12,13] Such methods have two major advantages: they are applicable to compounds of diverse structural type as well as to series of structurally similar compounds, and they can be used with large computer-based files containing chemical structures and property data,[14] with automatic derivation of appropriate structural features from computer-readable structural representations.[15] Thus, a recent example of substructural analysis[13c] gave simulated property predictions for a group of structurally diverse local anaesthetics by regression analysis, using structural features automatically derived from connection tables. As yet, however, no systematic method of general applicability has been developed for representing positional isomerism adequately for such analyses. Structural features derived from Wiswesser Line Notation (WLN) and used for structure–property correlation have included substituent patterns as part of the features representing ring systems,[13d] while structural features automatically derived from connection tables for application in information retrieval have included substituents as part of ring system features.[16] A recent study of boiling point variation within homologous series[17] used structural features which allowed for steric and dipolar interactions between identical substituents on a ring system, and obtained very high correlations which enabled simulated assignment of unknown stereochemistries, showing that chemical insight may be gained from such empirical methods.

It is evident that the ability to derive structural features so as to take account both of the position of a substituent relative

* To whom correspondence should be addressed.