in heterocyclic chemistry improves, it will be possible to formulate more precise and rigorous rules.

ACKNOWLEDGMENT

Gratitude is expressed to the National Science Foundation and Research Corporation for support of this work. The assistance of Timothy D. Salatin in developing the CAMEO program is also gratefully acknowledged.

REFERENCES AND NOTES

(1) Camille and Henry Dreyfus Foundation Teacher-Scholar, 1978-1983; Alfred P. Sloan Foundation fellow, 1979-1981.
(2) (a) Salatin, T. D.; Jorgensen, W. L. *J. Org. Chem.* **1980**, *45*, 2043. (b) Corey, E. J.; Wipke, W. T. *Science (Washington, D.C.)* **1969**, *166*, 178. (c) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. *J. Am. Chem. Soc.* **1972**, *94*, 421, 431. (d) Wipke, W. T.; Dyott, T. M. *Ibid.* **1974**, *96*, 4825, 4834. (e) Corey, E. J.; Petersson, G. A. *Ibid.* **1972**, *94*, 460.
(3) Frérejacque, M. *Bull. Soc. Chim. Fr.* **1939**, *6*, 1008-1011.
(4) Gasteiger, J.; Jochim, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43-48.
(5) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60-63.
(6) Wipke, W. T.; Dyott, T. M. *J.* Chem. *Inf. Comput. Sci.* **1975**, *15*, 140-147.
(7) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40-43.
(8) Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204-206.
(9) Paton, K. *Commun. ACM* **1969**, *12*, 594-598.
(10) Gibbs, N. E. *J. Assoc. Comput. Mach.* **1969**, *16*, 564-568.
(11) Dewar, M. J. S.; Gleicher, P. *J. Am. Chem. Soc.* **1965**, *87*, 685.
(12) Paquette, L. A. "Principles of Modern Heterocyclic Chemistry"; W. A. Benjamin: New York, 1968; pp 184-185.
(13) Forsen, S.; Wilsson, M. "The Chemistry of the Carbonyl Group"; J. Zabicky, Ed., Interscience: New York, 1970; Vol. 2, p 135.
(14) Garratt, P. J. "Aromaticity"; McGraw-Hill: London, **1971**; p 131.
(15) Gilman, H.; Melstrom, D. S. *J. Am. Chem. Soc.* **1948**, *70*, 1655.
(16) Adams, A.; Slack, R. *J. Chem. Soc.* **1959**, 3061; *Chem. Ind. (London)* **1956**, 1232.
(17) Reference 21, p 135.
(18) Reference 13, p 167-183.
(19) Cook, M. J.; Katritzky, A. R.; Linda, P.; Tack, R. D. *J. Chem. Soc. Perkin Trans. 2* **1972**, 1295.
(20) Beak, P.; Covington, J. B. *J. Am. Chem. Soc.* **1978**, *100*, 3961.
(21) Katritzky, A. R.; Boulton, A. J., Eds., "Tautomerism of Heterocycles"; Academic Press: New York, 1976; pp 204-212, 249-250, 388-389, 414-415, 444-445.
(22) Katritzky, A. R. *Chimia* **1970**, *24*, 134.
(23) Reference 12, p 205.
(24) Frank, J.; Katritzky, A. R. J. *Chem.* Soc. *Perkin Trans.* 2 1976, 1428.
(25) Reference 13, pp 195-197.
(26) Reference 21, pp 388-389.
(27) Baba, H.; Takemura, T. *Bull. Chem. Soc. Jpn.* **1964**, *37*, 124. Fieser, L. F. *J. Am. Chem. Soc.* **1931**, *53*, 2329. Clar, E. *Chem. Ber.* **1949**, *82*, 495.
(28) Hornfeldt, A. B. *Arkiv. Kemi* **1968**, *29*, 461.
(29) Reference 21, pp 405-415.
(30) Tanaka, T. *J. Pharm. Soc. Jpn.* **1971**, *91*, 338.
(31) Reference 21, 443-445.
(32) Reference 21, pp 303-304.
(33) This prevents tautomerizing functional groups such as oximes.
(34) Schulenberg, J. W. *J. Am. Chem. Soc.* **1968**, *90*, 7008.
(35) Mazharuddin, M.; Thyagarajan, G. *Tetrahedron* **1969**, *25*, 517.
(36) Mason, S. F. *J. Chem. Soc.* **1954**, 2071.
(37) Lister, J. H. *Adv. Heterocycl. Chem.* **1966**, *6*, 1.
(38) Pal, B. C.; Horton, C. A. *J. Chem. Soc.* **1964**, 400.
(39) Cavalier, L. F.; Fox, J. J.; Stone, A.; Change, N. *J. Am. Chem. Soc.* **1954**, *76*, 1119.
(40) Angell, C. L. *J. Chem. Soc.* **1951**, 504.
(41) Bugg, C. E.; Thewalt, U. *J. Am. Chem. Soc.* **1970**, *93*, 7441.
(42) Reference 21, pp 211, 253, 484-485.
(43) Reference 21, p 250.

CBF—Computer Handling of Chemical and Biological Facts. 2[1]

J. BECKER and D. JUNG*

Electronic Data Processing Department, Boehringer Ingelheim KG, 6507 Ingelheim am Rhein, Federal Republic of Germany

W. KALBFLEISCH and G. OHNACKER

Research Administration Department, Dr. Karl Thomae GmbH, 7950 Biberach an der Riss, Federal Republic of Germany

Received July 29, 1980

CBF[1] is an EDP-supported documentation and retrieval system for structural formulas of defined organic compounds and their biological activities. It fits the needs of a firm concerned with drug research for prevailing unpublished internal data, which the scientists can use for reflections on structure-activity relationships and to search for lead compounds with special activity profiles.

INTRODUCTION

The CBF system has been used successfully for 12 years by various research centers of C. H. Boehringer Sohn, Ingelheim. It was conceived as a data input system for

(1) storing biological screening results from drug research in a computerized data base,

(2) providing printed information about chemical compounds and/or screening results either as a continous service in file-card form or as printouts of results of retrospective searches.

Our files contain connectivity tables of 170 000 chemical structural formulas as well as 260 000 individual results from biological screening tests of 78 000 substances. As could be expected, everyday use of this system over the years has illuminated several features requiring improvement. Innovations were undertaken with respect to the program, and certain alterations to individual elements of the entire system were introduced. The aim of such improvements was to rationalize data input and to expand search capabilities.

INPUT OF CHEMICAL INFORMATION

We are using the methods of machine transformation into a condensed connectivity table to store all structural information unambiguously.[2,3] Together with the connectivity table, a series of screens is machine generated. Structural formulas requiring maximally 256 nodes (nonhydrogen atoms) in the topological list can be currently handled.

As a consequence of topological storage, only unambiguously defined structures can be stored in a retrievable form. Substances whose structures are equivocal can be stored either in

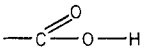
0095-2338/81/1621-0111$01.25/0

the name or print format; this storage form naturally cannot be retrieved in a chemical search but can be included in the printout of biological searches.

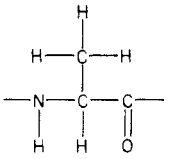
For formula input optimization, a microprocessor-controlled semigraphic device was developed. It has proved effective in daily use over the past three years[4] and is as effective as other input systems described in the literature. Its remarkable advantage is the favorable price/performance relationship.

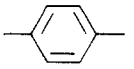
For simplication of chemical input and presentation of a formula display familiar to the chemists, abbreviations for substructures with one or two linkage positions can be employed. These abbreviations can be defined by the user. They may consist of one to eight characters. Their connectivity tables are stored in a special file from which they are incorporated into the appropriate structure positions. Besides these stored abbreviations there is the possibility of using abbreviations for unbranched chains of the form C_nH_{2n+1} or $-(CH_2)_n-$ with $1 \leq n \leq 99$. The connectivity tables for these substructures are generated during the input.

Examples:

abbreviations
–COOH instead of 

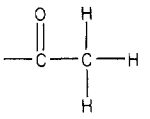
$-NO_2$ instead of 

–ALA– instead of 

–P-PHENYL instead of 

$-(CH_2)_5-$ instead of

For substructure abbreviations synonyms can be defined, e.g., $-COCH_3$, CH_3CO-, and ACETYL can be used instead of



For a minimization of time-consuming searches in the topological matrices, the following screens are generated by our CBF program during the input process:

numerical screens

basic screen	number of atoms, heteroatoms, rings, heterorings, bonds, and modifications
empirical formula screen	type and number of atoms
modification screen	quantity of additional stored information (e.g., positive charge, cis form, radioactive labeling, information about steric arrangement, etc.)

bit screens

fragment screen	presence of particular structural fragments, e.g., —COOR, $—CONH_2$, —C≡N, $—N^+R_4$, —CO—O—CO—, —N—N=R_2
ring screen	code describing types and linkages of rings, which are contained in the structure.

The effectiveness of these screens and an optimal search strategy is dealt with in the Technical Data section.

The CBF system provides the possibility to print the stored chemical structures. For this the structural and other chemical information is stored as drawn on the screen.[4] Our IBM type printer chain unit equipped with special chemical characters allows printing of formulas familiar to chemists (see Figure 1).

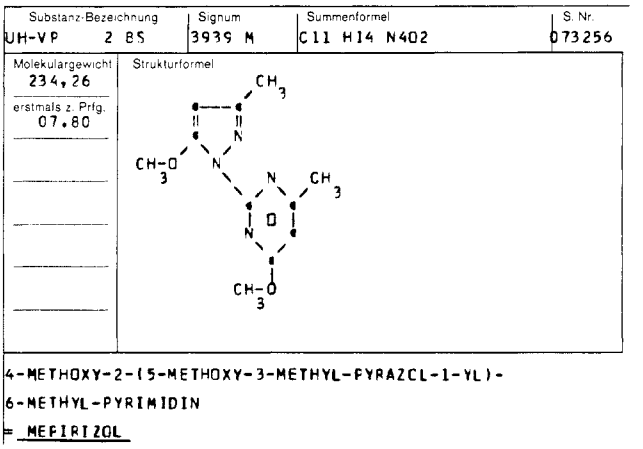


Figure 1. Printout of a chemical formula on a chemical filing card.

INPUT OF BIOLOGICAL INFORMATION

A "general format" was developed for the input of different biological information (see Figure 2). It comprises four sections: (1) the "test heading" deals with basic information—as part of the test heading, for each single test, a test-modification number and the date of testing is stored; with these means, precautions are taken to find out data for comparisons of biological activities, which are in fact comparable; (2) the "observation section" for observed or measured dose-dependent values; (3) the "result section" for results, which can be calculated from the values of the observation section, e.g., ED_{50}, LD_{50}; (4) a nonsearchable commentary for brief evaluation.

The various test methods are distinguished unequivocally by test numbers. If required, the dose dependent values and results can be specially characterized by descriptors. Furthermore, special codes are introduced for animal species, application modes, observed side effects, and units of measurements. The observation section can also include several individual measurements per dose level (e.g., if the measurements are taken at different times). Correspondingly, summarized results may also be included in the results section (e.g., ED values at different times).

The general format has a modular construction; its individual modules can be arbitrarily composed to "specialized formats" for the coding of particular, standardized tests (e.g., Figure 3).

PRODUCTION OF INDEX CARDS

Besides its purpose as a retrieval system, CBF provides the production of chemical and biological index cards (see Figures 1 and 4). They are generated either synchronously with each data input or by special order. For the biological card file a table controlled and format specific printout processor is used. The number of printout copies can be chosen as required. Optionally the generated card files can be sorted according to empirical formulas, internal substance code, molecular weight, or date of input. Naturally, these sorting terms can also be combined. This service provides the different users with standardized information and supports the scientists in developing individual card files.

RETRIEVAL

Questions can be formulated for searches in the chemical and/or biological file. Within the searches Boolean logic can

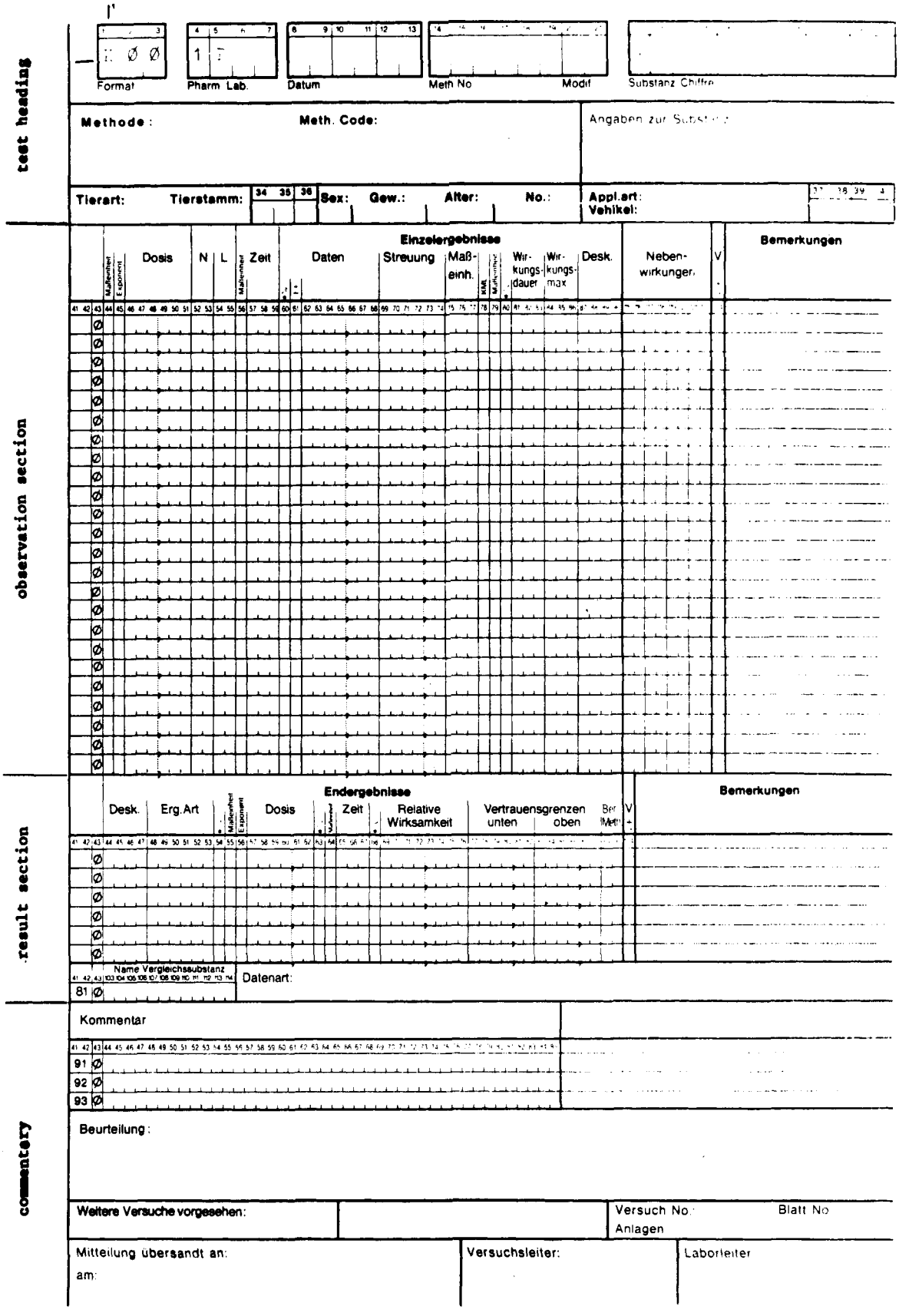


Figure 2. Data sheet for the general format.

be used. For a single batch run up to 50 individual search profiles can be pooled.

Chemical Searches. Within the chemical file searches can be done for definite structures and for parts of structures as well as for classes of compounds and structures represented by general formulas (Markush formulas).

Substructure Searches and Partial Structure Identity. Our program enables us to carry out in one profile topological searches of maximally 26 different substructures within one stored structure. Furthermore, it is possible to define sub-

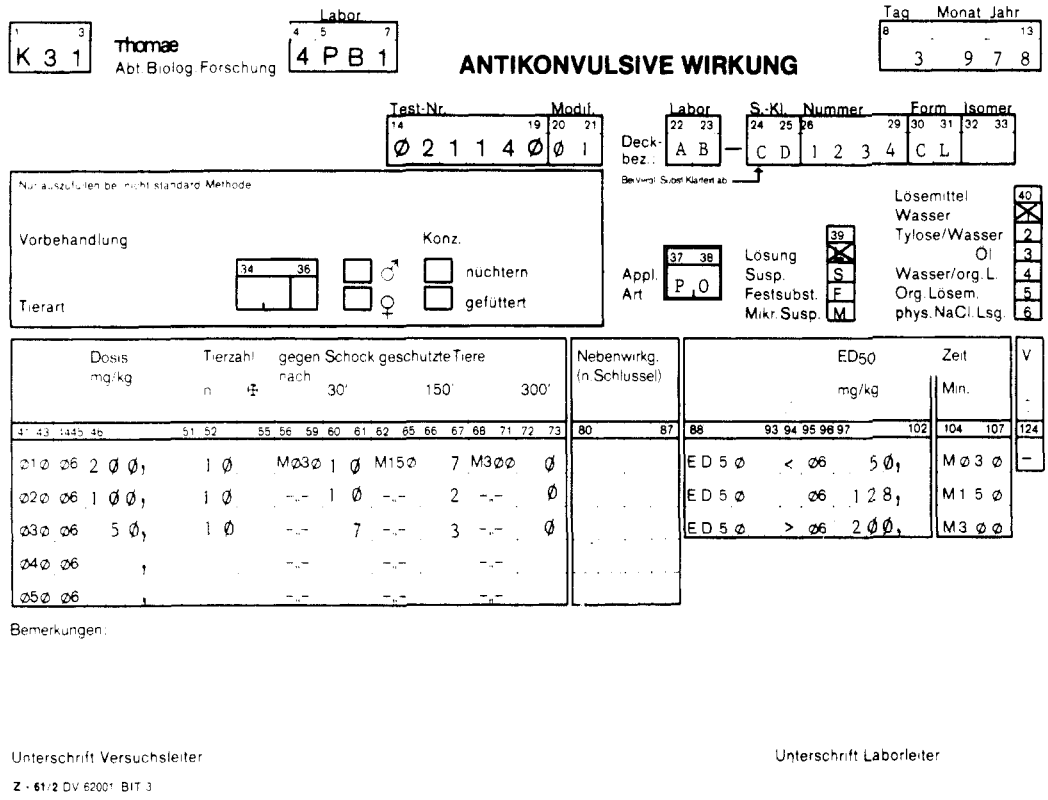


Figure 3. Special data sheet for anticonvulsive activity.

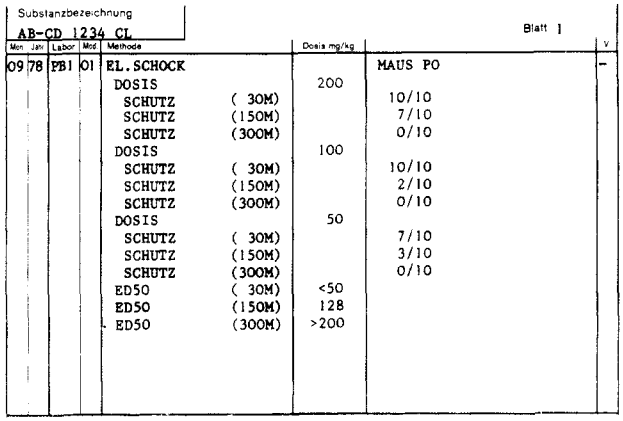
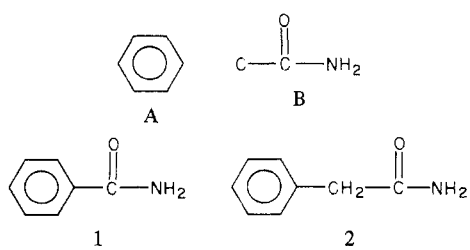


Figure 4. Biological filing card. The printout corresponds to the data entered in Figure 3.

structures, for which positive or negative search logic is applied, so that they possess or do not possess at least one common atom, i.e., have partial identity or not.

Example: The structures A and B are substructures of both the stored structure 1 and the stored structure 2. In structure 1 they have partial identity but not, however, in structure 2.

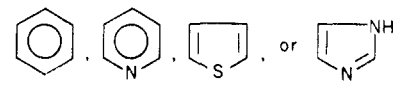


Partial identity can be stipulated or forbidden for any substructure either for an arbitrary or a specified atom. Partial identity operations are not limited to individual pairs of substructures but may also be formulated at any required depth of logic, i.e., the inclusion or exclusion of partial identity of substructures, which are themselves partially identical with further substructures and so on, may be stipulated.

Finally, a list of substructures can be defined, from which at least one member is partially identical with another substructure or any member of a further list of substructures. It can, for instance, be stipulated that one of the substructures

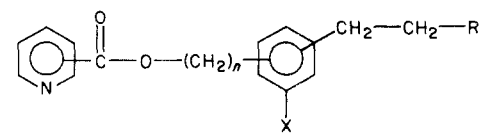
$$C\text{–}Cl,\ C\text{–}CF_3,\ C\text{–}NO_2,\ \text{or}\ C\text{–}OCH_3$$

is partially identical with one of the substructures



That means, any of the rings above can be substituted at any C atom with one of the functional groups –Cl, $-CF_3$, $-NO_2$, or $-OCH_3$.

Substructure searches and partial identity operations can also be combined without restrictions, employing logical operators. These operations enable the user to carry out searches for defined groups of compounds which can be summarized as Markush formulas, e.g.,



with n = 1–3, X = Cl, Br, CF_3, or NO_2, R = arbitrary substituent, but not $N(CH_3)_2$. The ring can be substituted at any position.

Biological Searches. The search modes in the biological file are directed to searches for particular biological activity profiles, whereby individual information elements are invoked by means of keywords. Stipulations above or below particular activity levels may be made, definite side effects may be re-

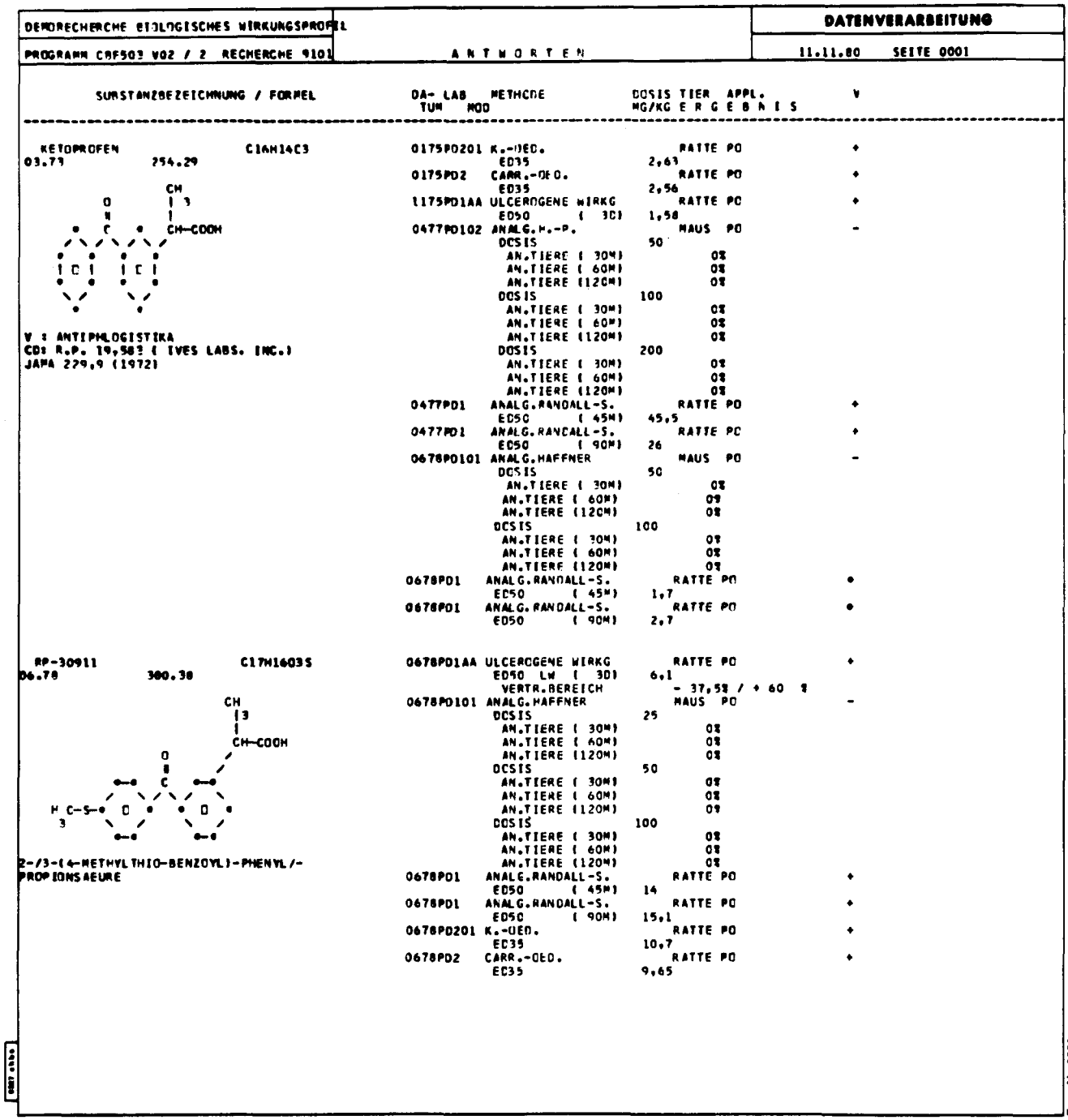


Figure 5. Printout of demonstration search 1.

quested or excluded, etc. In the construction of the required activity profile, free Boolean logic combinations of the search terms are allowed. These options provide the user with very distinct information which is free of ballast.

Because our substance code is hierarchically constructed (the last 4 characters of a 12-character string describe the salt form of a compound), searches for activity profiles are feasible with the CBF system, even if the individual tests from which the desired profile is compiled have been carried out with different salt forms of the identical basic structure.

Printout Processing. In the case of a hit, in the first step the total information (chemical and biological sections) concerning the hit compound is made available for an editing post-processor program which allows the selection of particular biological results as well as a highly flexible listing format for the printout of search results. The list formats for the printout may be generated in different ways on demand. Beside the two fixed forms (see Figures 5 and 6), an arbitrary variable format can be generated.

SEARCH EXAMPLES

Demonstration Search 1 (Printout; See Figure 5): Substances are searched by using those described in the literature and stored in our file, with the following activity profile:

- an ED_{35} value of under 40 mg/kg (po), using the kaolin- or carrageenin-edema test in rats
- an ED_{50} of more than 1 mg/kg (po), for the ulcerogenic activity in rats
- an analgesic test according to Randall-Selitto, an ED_{50} of less than 100 mg/kg at 45 min and less than 50 mg/kg at 90 min after administration of the test compound are requested
- analgesic tests, according to Haffner and the Hotplate method, should exhibit no activity.

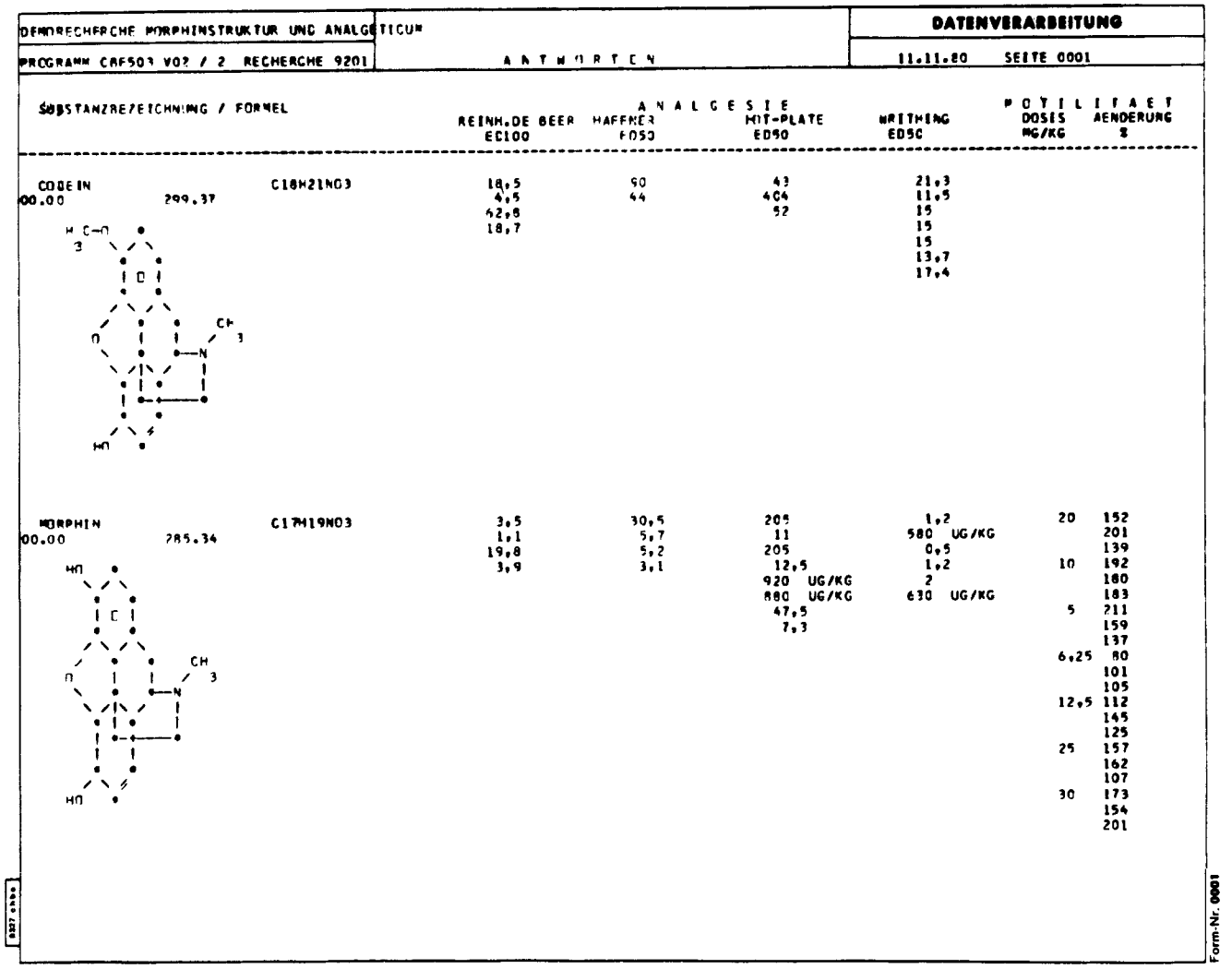


Figure 6. Printout of demonstration search 2.

The test results shall be printed in the index card format.

Demonstration Search 2 (Printout; See Figure 6): Substances are searched with the morphine ring system and an ED_{50} of at least 0.5 mg/kg (sc) in the Hotplate test on mice. In addition to the chemical structure, the ED values of all analgesic tests should be printed out in columns, ordered according to the individual methods. The changes in the motoricity following various doses should also be given.

TECHNICAL DATA

Program language

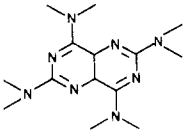
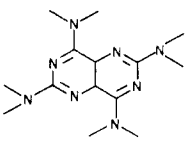
Program language

program series	program language	program size, K bytes
CBF 100 (chemistry input)	PL/1	296
CBF 200 (biology input)	PL/1	186
CBF 500 (retrieval)	assembler, PL/1	300

Time Requirement for Searches. The CPU times given were measured on an IBM 370/158 computer.

Case A. Substructures containing chains of at least 10 C atoms and a terminating chlorine atom are sought (see Table I). Searches 2 and 3 show that CPU time for topological searches is drastically reduced by optimal coding. As one can see it is of high importance to start the numbering at a less common atom.

Table I. Search Time on Substructures

search no.	coding strategy	mode of search	hits	CPU time, min
		Case A		
1	C_1-C_2-C_3···C_{10}-X_{11} X = arbitrary atom	topological searches *without* screen	38368	35.4
2	C_1-C_2-C_3···C_{10}-Cl_{11}		3354	88.5
3	Cl_1-C_2···C_{11}		3354	3.2
4	Cl > 0, C > 9	empirical formula, screen *only*	24392	1.3
5	combination 3 + 4	empirical formula, screen and topology	3354	2.1
		Case B		
6		topological search *without* screen	356	6.0
7		*with* screen, applying ring screen and empirical formula screen	356	0.9

Case B. 2,4,6,8-tetraaminopyrimido[5,4-*d*]pyrimidines were sought, whereby the amino groups are substituted.

Table II. Search Time on Biological Activity Profiles

	substances qualified as hits	CPU time, min
demonstration search 1 (see Figure 5)	25	2.3

For search time on biological activity profiles, see Table II.

CONCLUSION

(1) CBF is an easy to handle system for storing and retrieval of structural formulas and results of biological tests.

(2) The biological section of CBF has a modular construction. Therefore, new biological tests can be adapted easily and quickly to this section. It can be employed also in the documentation and retrieval of other properties, such as analytical or physicochemical data.

(3) CBF was originally developed as an offline system one-and-a-half decades ago. Meanwhile, the users demand more and more direct and quick access. Therefore, we are working at present on methods to bring CBF data online to the scientists bench.

REFERENCES AND NOTES

(1) G. Ohnacker, W. Kalbfleisch, "CCBF—A System for the Computer Processing of Chemical and Biological Facts", *Angew Chem., Int. Ed. Engl.*, **9**, 605-610 (1970). (The earlier designation "CCBF" is now abbreviated to "CBF".)
(2) D. J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont", *J. Chem. Doc.*, **5**, 43-51 (1965).
(3) H. L. Morgan, "The Generation of A Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107-113 (1965).
(4) W. Kalbfleisch, G. Ohnacker, "Off-Line Input of Chemical Structures with a Low-Cost Microprocessor-Controlled Semigraphical CRT", *J. Chem. Inf. Comput. Sci.*, **20**, 176-178 (1980).

NEWS AND NOTES

CAS SENIOR EXECUTIVES

Three senior executives of the American Chemical Society's Chemical Abstracts Service (CAS) have been given new responsibilities.

CAS Director Dale B. Baker has been named chief operating officer for American Chemical Society Columbus operations. He also will continue to serve as director of CAS, a post he has held since 1958.

Ralph E. O'Dette, senior staff advisor to the director of CAS, has been given additional responsibilities as director of planning. In this capacity, he will lead the organization's short- and long-range planning efforts. O'Dette has been senior staff advisor since 1966.

James V. Seals, Jr., has been appointed to the new post of director of international programs with responsibility for CAS's cooperative efforts with organizations in Europe and Japan and its other international activities. Seals has been in West Germany as manager of international markets for CAS for the past two years. Prior to that, he served as international liaison officer for Internationale Dokumentationsgesellschaft fuer Chemie in West Germany and in a number of posts at CAS, including manager of the special services department and assistant to the director of research and development.

CAS SEARCH ASSISTANCE

The Chemical Abstracts Service Search Assistance Desk can now be reached through a toll-free "800" number—(800) 848-6533. The new number serves searchers throughout the United States with the exception of calls placed in Ohio. Residents of Ohio should continue to call (614) 421-6940, ext. 3209.

The desk is staffed by chemists who will answer questions about how CAS indexing terminology and practices affect searches for references on particular topics or substances in *Chemical Abstracts*, the *CA Search* computer-readable file (including online chemical dictionary files), and the new direct online chemical information service, CAS ONLINE.

The Search Assistance Desk is in operation Monday through Friday from 8 a.m. to 5 p.m. Eastern time. Written questions may be directed to Search Assistance Desk, Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210.

NEW CAS SERVICES

Chemical Abstracts Service (CAS) is offering two additional services based on its Chemical Registry System computer file. In one, the Private Registry Service, CAS will build computer files of substance information from an organization's private files, maintain them on CAS computers, and provide remote online access to the files for searching. In the other, the Registry Profile Service, individuals or organizations can request information on specific chemical substances from the CAS Chemical Registry System in a variety of ways and, if desired, receive notification whenever updated information on a substance of interest enters the system.

The Registry Profile and Private Registry Services complement the CAS ONLINE service, introduced last year, which provides online access to information in the CAS Chemical Registry System through remote terminals.

Through the Private Registry Service, CAS will create and maintain a computer file of structural and other information on substances in a company's or organization's private files. The organization's employees will be able to search the private file remotely, using CAS ONLINE search techniques. These techniques make it possible to search for substances that share structural features and display or printout structure diagrams of the answers as part of the search results. The organization's employees will also be able to update and retrieve other information in the private file, search substance information from the CAS Registry file through CAS ONLINE to determine if a substance of interest has been reported in the literature, and periodically check new information entering the Registry to discover when a substance first appears in the literature. Fees for the Private Registry Service depend on the size, content, and form of the organization's substance file and kind and frequency of searches performed.

The Registry Profile Service enables individuals or organizations to obtain Registry information about specific substances of interest. Requestors can identify substances by CAS Registry Numbers, names, structure diagrams, or Wiswesser Line Notations. Information retrieved can include CAS Registry Numbers, *Chemical Abstracts* index names, molecular formulas, and structure diagrams. The retrieved information will be provided in printed or computer-readable form at the requestor's option.

Other options available in the Registry Profile Service include searches of the Registry file, using CAS ONLINE search techniques to identify substances that contain particular structural features of interest (substructure searches), periodic current-awareness searches of the file, and automatic forwarding of new information added to the Registry file on specific substances of interest. Common and trade names on file for substances and the computer-generated connection table records that describe structures of substances in the Registry System can also be obtained through the service but are subject to use restrictions. Fees for the Registry Profile Service depend on the form in which substances are identified in the request, the information requested, and the use for which