of the Agency, and no official endorsement should be inferred. The PRIME 750 computer used in these studies was purchased, in part, with the support of the National Science Foundation.

REFERENCES AND NOTES

(1) Nilsson, N. J. "Learning Machines"; McGraw-Hill: New York, 1965.
(2) Tou, J. T.; Gonzalez, R. C. "Pattern Recognition Principles"; Addison-Wesley: Reading, MA, 1974.
(3) Stuper, A. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1976**, *4*, 238.
(4) Schrage, Linus *Assoc. Comput. Mach. Trans. Math. Software* **1979**, *5*, 132.
(5) Muller, Mervin E. *J. Assoc. Comput. Mach.* **1959**, *July*, 376.
(6) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley-Interscience: New York, 1979.
(7) Moriguchi, Ikuo; Komatsu, Katsuichiro; Matsushita, Yasuo *J. Med. Chem.* **1980**, *23*, 20.
(8) Pietrantonio, Lucio; Jurs, Peter C. *Pattern Recognition* **1972**, *4*, 391.
(9) Fleiss, J. L. "Statistical Methods for Rates and Proportions"; Wiley: New York, 1973.
(10) Whalen-Pedersen, E. K.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 264.
(11) Topliss, John G.; Edwards, Robert P. *J. Med. Chem.* **1979**, *22*, 1238-1244.

Substructure Searching of Heterocycles by Computer Generation of Potential Aliphatic Precursors

RICHARD L. M. SYNGE

School of Chemical Sciences, University of East Anglia, Norwich, NR4 7TJ, England

Received May 8, 1984

Heterocyclic structures, in natural and synthetic organic compounds, are mostly formed by cyclizations of aliphatic precursors. Computer programs have been written that notionally reverse this process. By breaking, in turn, one of the heterobonds in each ring of each heterocyclic region of a molecule, a permuted set of aliphatic tree structures is generated. Exhaustive search of the "bonded-atom" strings present in such a set of trees can reveal unsuspected structural and biosynthetic aspects of a molecule. Some current methods of substructure searching, by over-emphasizing cyclic structures, fail to detect these relationships.

INTRODUCTION

Heterocyclic structures tend to arise, both in nature and in the laboratory, by cyclizations of aliphatic compounds. It was recognition of this general principle that gave Robert Robinson such great insight into structures and laboratory syntheses of natural products.[1]

Much of what Robinson grasped intuitively has since been confirmed (in a general way) by biosynthetic experiments using isotopic tracers. Conversely (and using computer programs rather than experiment), it should be possible to "disinter", from heterocyclic structures, some of their hypothetical aliphatic precursors, thus throwing light on analogies in the biosynthetic pathways that have produced them and, perhaps, on their pharmacological properties. This paper describes one possible approach to this goal. This may also turn out helpful as a cheap (and, in some respects, impressionistic) aid in substructure searches of not-too-large files. It may also have its uses in tracking down compounds that fall within "generic" (Markush) claims in chemical patents.

In general, chemical nomenclatures and notations have been based, almost obsessively, on cyclic structures. This has been an inevitable consequence of the classificatory system developed in *Beilstein*. It is interesting that Prager and Jacobson, in their preface to the still-current 4th edition,[2] emphasized that *different* systematic arrangements were needed for particular purposes. Such a need is manifest in terpenoid chemistry, where the cyclizations are of less interest than the aliphatic skeletons that may or may not become cyclized. Randić and Wilkins[3] have developed interesting algorithms for searching such structures. Their approach, like that of this work, attaches equal emphasis both to cyclic and to aliphatic structures.

Graph theory[4] shows that, if *one* interatomic bond in each ring of a cyclic compound is broken, an aliphatic tree structure will result, without disconnection of any part of the molecule. In devising the present approach, heterocyclic moieties (or "domains") of each molecule have been treated separately, where a carbocyclic structure comes between them, and carbocyclic structures have been excluded from the manipulations. In each heterocyclic "domain", the ring bonds in which heteroatoms participate are broken in turn, so that all possible break permutations give rise to aliphatic trees (see Figure 1).

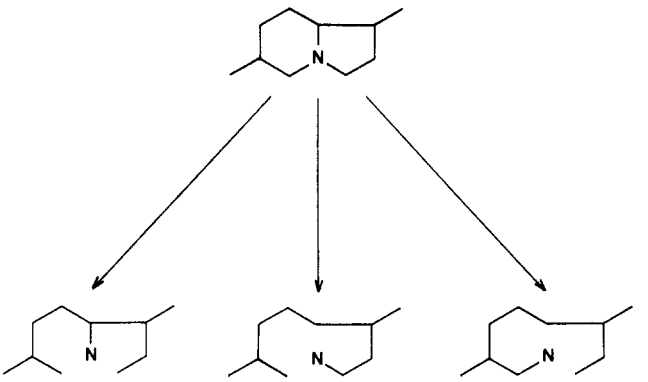


Figure 1. Formulas (lower) show the three heterobond break permutations by which the (upper) bicyclic heterocyclic compound is aliphatized.

Heterobond breakage in symmetrical compounds will yield the same aliphatic product more than once. However, such symmetry is rare among organic compounds, so nothing has been done to curtail resulting redundancies. Stereochemistry has been ignored throughout. Bond breakage is notionally by hydrogenolysis—thus, the structures produced in Figure 1 are all aliphatic primary amines. Notionally too, unsaturated

0095-2338/85/1625-0050$01.50/0

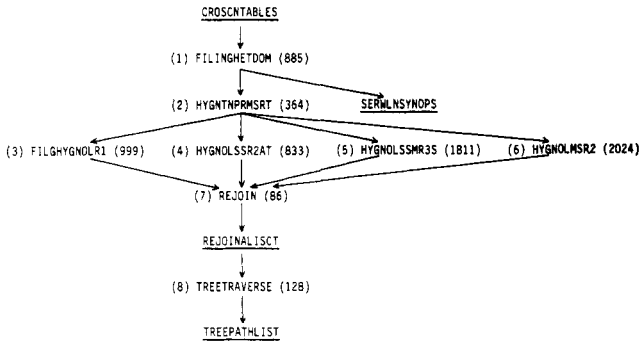


Figure 2. Flowsheet for the eight programs (appended numbers give lines of FORTRAN-77 in program). File names for input/output files mentioned in text are underlined.

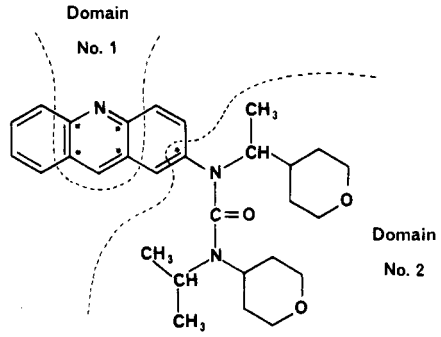


Figure 3. Subdivision of a compound by program 1 into two "heterocyclic domains".

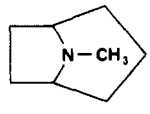
compounds have initially been hydrogenated; this not only eliminates ambiguities arising from tautomerism but also has considerable biological justification and helps to juxtapose analogous structures.

The essential feature of the rearrangements is that all C–C bonds are left intact since, in biosynthetic processes, it is usually simple aliphatic carbon skeletons that become substituted with heteroatoms, either before or in the course of heterocycle formation.

"CROSSBOW" connection tables[5] have been used as initial data for the structural manipulations. These have the particular advantage that they specify explicitly each ring of the "smallest set of shortest rings".[6] There is the additional advantage that computer programs are commercially available[7] for generating CROSSBOW connection tables from files of WLN formulas, such as those of new compounds published by *Current Abstracts of Chemistry & Index Chemicus* in connection with *Chemical Substructure Index*.

SCOPE AND LIMITATIONS OF THE PROGRAMS

The programs described in this paper are written in FORTRAN-77. They deal with compounds containing simple and/or fused rings. External-spiro and ring-of-rings compounds have not been handled, although the algorithm for aliphatizing these is identical. For bridged-ring compounds, only one bond in a bridge may be broken, in any break permutation, else disconnection ensues. In consequence, some bridged-ring etc. compounds such as tropane (**1**) would not be usefully handled, since a rather "meaningless" carbocyclic moiety results from breaking either of the heterobonds of such



1

a ring system. For "perifused" compounds, additional restrictions apply, so that atoms at "multicyclic points" do not become disconnected. Modified algorithms, to deal with all these categories, would be easy to devise.

Subject to these limitations, the programs, as now written, handle molecules initially containing up to 59 non H atoms (C, N, O, P, S, and halogens only), up to 20 rings (each having up to nine ring atoms), up to 20 ring systems, and up to 12 "connection transfers".[5] The flowsheet (from initial data CROSCNTABLES to final output TREEPATHLIST) is given in Figure 2.

DESCRIPTION OF ALIPHATIZATION PROGRAMS

Program 1 processes initial data into separate CROSSBOW connection tables for each heterocyclic domain. All C atoms that have come from any carbocyclic ring are specially labeled (see starred atoms in Figure 3), and this labeling is retained through all the later programs. At the same time, the program produces a file summarizing, for each compound, the rings and ring systems in the molecule and the "domains" into which it has been divided. For the sample molecule in Figure 3, the three-record entry in SERWLNSYNOPS is shown in Figure 4.

Program 2 elminates unwanted unsaturations by altering some of the "bonded-atom" characters.[5] As the program stands, most unsaturations in rings or chains are removed by notional hydrogenation; most other functional groups, whether attached to rings or not, are not modified, although imines are hydrogenated to amines and oximes to hydroxylamines. The program then specifies the permutations of ring breakages (upper limit 98 permutations). It then sorts the processed connection tables (with break permutations) into four files: (a) single rings; (b) pairs of fused rings; (c) fused-ring systems (up to five rings); (d) domains including more than one ring system (up to five rings in all, with not more than two in any one ring system).

Programs 3–6. Each of these works on its corresponding input file to effect treeing by hydrogenolysis. There is further reductive modification of some of the bonded-atom symbols for those atoms involved in bond breaking. Output is to four respective files each having four-record entries, in identical format, for each break permutation handled (Figure 5).

Program 7 simply serializes the entries in the above four files into a single sequence according to (a) serial number of compound, (b) heterocyclic domain number, and (c) serial number of break permutation.

SEARCHING THE ALIPHATIZED STRUCTURES

It would be possible to generate, from the connection tables in REJOINALISCT (see Figures 2 and 5), readable, though uncanonical, WLN, much as described by Lynch.[8] However, the hydrogenolysis programs (3–6) give rise to untidily numbered trees (see Figure 5), which require quite long programs to reduce them to WLN-like order. Moreover, canonical or not, branched aliphatic WLN is particularly unsuited for string searching,[9] and excessively complicated search programs would be needed to avoid an unacceptable proportion of misses.

Accordingly, the bonded-atom symbols have been left intact in their connection tables (Figure 5) and a simple algorithm[10] was used (program 8, TREETRAVERSE, Figure 2) for listing the unique paths from each branch end to the root and then from each branch end in turn to each remaining branch end. Figure 6 shows the resulting paths for a small structure having two "connection transfers".[5] Structures with 10 connection transfers (12 terminal non H atoms, including the root) generate 66 such paths. (Break permutations having >10 con-

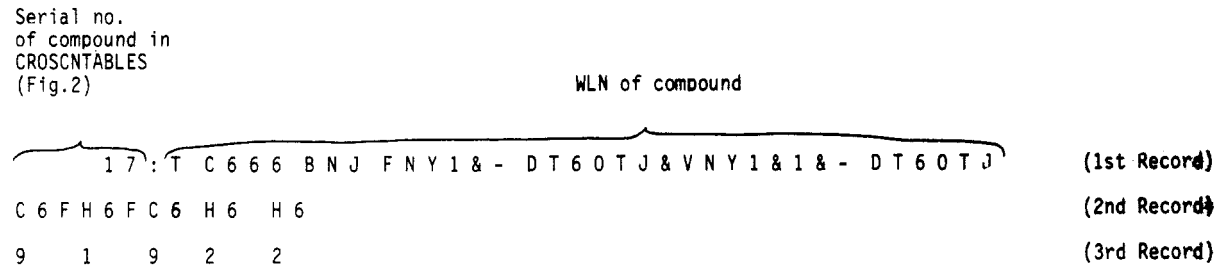


Figure 4. Three-record entry in SERWLNSYNOPS for the compound shown in Figure 3. In the second record, "C" signifies a carbocyclic and "H" a heterocyclic ring, followed by the number of ring atoms and preceded by "F" if fused to preceding ring. In the third record are given the domain numbers to which each ring has been assigned (eliminated carbocycles are labeled "9"). If only heterocycles occur in the molecule, all are labeled "0", in which case the entire connection table is passed intact to next program (Figure 2).

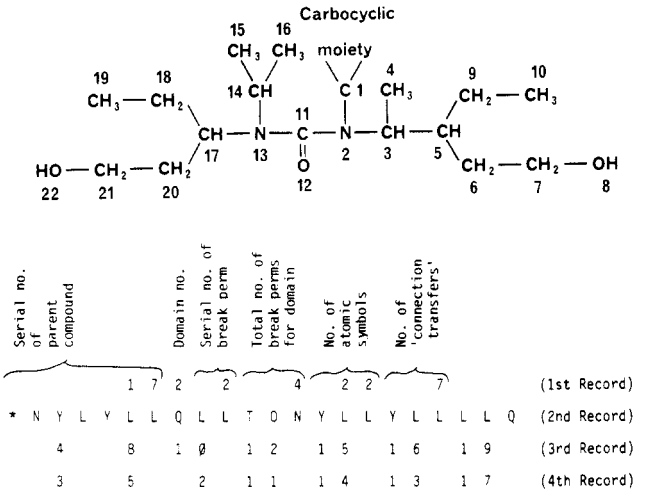


Figure 5. (Top) Aliphatic structure resulting from one of the break permutations in domain 2 of the compound illustrated in Figures 3 and 4. Numbering of non H atoms is as in the connection table beneath. (Bottom) Four-record entry in REJOINALISCT for same. The significance of the six numbers in the first record is annotated; the second record gives the 22 bonded-atom symbols in the sequence shown in the formula. The integers in the third record show connectivity "ends" (terminal atoms), and those in the fourth record show corresponding atoms from which connectivity is resumed.

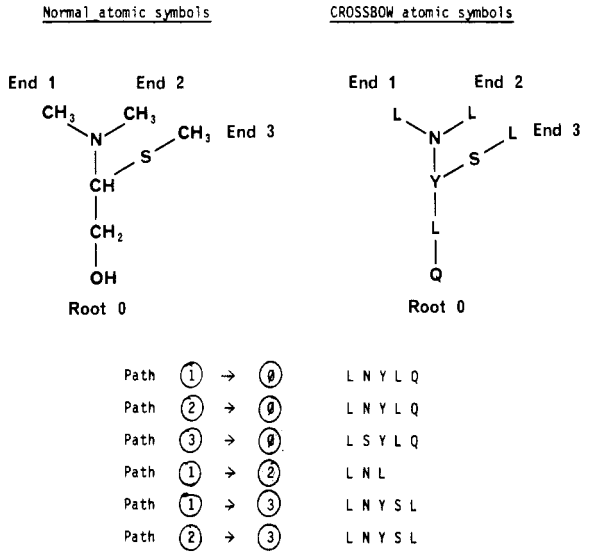


Figure 6. Comprehensive listing of tree paths for an aliphatic structure having two connection transfers.

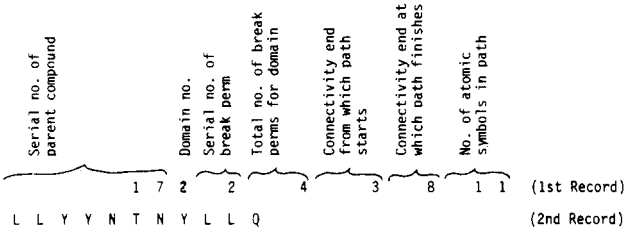


Figure 7. Two-record entry in TREEPATHLIST relating to the break permutation illustrated in Figure 5. The path runs from connectivity end 3 (atomic symbol 10) to the last symbol of the connection table in REJOINALISCT (atomic symbol 22). The significance of the seven numbers in the first record is annotated. The second record gives the 11 bonded-atom symbols in their sequence along the path.

nection transfers have been excluded from the final search file TREEPATHLIST generated by the program.)

With this algorithm,[10] every branch point in the structure is traversed at least once by every available route, so that no string of bonded atoms bonded to one another can be missed, wherever it occurs in the structure, provided that each of the paths is searched both forward and backward. A sample two-record entry in TREEPATHLIST, derived from the structure in Figure 5, is shown in Figure 7.

COMPOUNDS FOR TESTING

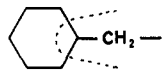
The initial data file (CROSCNTABLES, Figure 2) at present holds about 500 heterocyclic compounds. It started as a collection of ready-to-hand connection tables from a file belonging to Dr. Peter Willett, which was supplemented with special compounds to test particular sections of the various programs and with compounds having personal interest for the author. There were later added about 250 compounds randomly selected from *Chemical Substructure Index* (1971–1978). These were new compounds of greater-than-average complexity (due to the more frequent rotation of the longer WLN's in *Chemical Substructure Index*). There were finally added 80 eligible compounds selected at random from the *Merck Index*[11] and 80 such compounds selected at random from graphic formulas displayed in Volumes 14–21 of *Phytochemistry* (1975–1982).

The 500 compounds in CROSCNTABLES yielded 568 heterocyclic domains. A total of 3168 break permutations appeared in REJOINALISCT, an average of 5.6 permutations per heterocyclic domain. (Only the first 98 permutations were filed for any one domain.) A total of 53 534 atom paths (see Figure 7) occurs in the final search file TREEPATHLIST. (The exclusion from TREEPATHLIST of tables in REJOINALISCT having >10 connection transfers removed <1% of the permutations.) The final search file thus averages 16.9 atom paths per break permutation and 94.3 atom paths per heterocyclic domain. The average path length is 7.9 non H atoms.

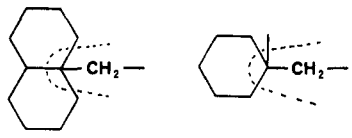
BONDED-ATOM SYMBOLS: SPECIAL CHARACTERS FOR CARBOCYCLIC C ATOMS

In general, the characters in current use in the CROSSBOW system have been taken over unchanged for the present work. However, three additional characters have been introduced to serve as labels for carbocyclic C atoms in a heterocyclic domain. These are "$" (replacing "X"), "*" (replacing "Y"), and ":" (replacing "L").

Where an isolated carbocyclic C atom is involved in a heterocyclic domain, that becomes "*" if in a carbocyclic periphery; thus

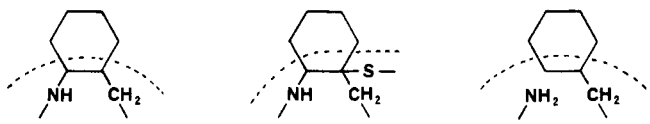


is represented "*L...". If it is at a carbocyclic fusion point or at a domain boundary, it becomes "$"; thus



are represented "$L...". Such isolated carbocyclic C atoms are always terminal in an atom path to be searched.

Where two adjacent carbocyclic C atoms are in a heterocyclic domain, the following codings are typical of what occurs:



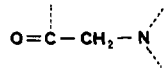
"...M**L..." "...M*$L..." "...M*$S..." ":*L..."

Here, "*" is never terminal and never at a branch point. "$" is always at a branch point included in the heterocyclic domain. ":" is always terminal and can only rise by hydrogenolytic breakage of the heterobond in which that carbocyclic C atom was originally involved.

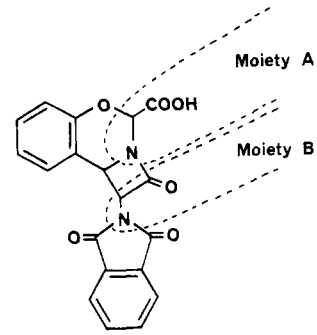
RESULTS OF STRING SEARCHES

Futher experience of these is needed, but early results are encouraging. All "hits" have contained the sought bonded-atom string. The absence of "misses", though much less easy to check, is equally demanded by graph theory.[10] It ought to be more selective, as well as saving computer time, if one end of the sought string is known to be terminal in the tree paths searched, and more so still if both ends are terminal [see (iv) below]. However, such restriction of string searches could lead to the missing of interesting structures in aliphatic side chains, in which heterobonds are not subjected to notional hydrogenolysis (*cf.* compound **14**). In the partial graphic formulas that follow, a dotted bond (···) indicates optional linkage to H or a non H atom, whereas an ordinary bond (—) indicates linkage to another non H atom. Wavy bonds (〰) are elements of carbocyclic systems. N atoms either are shown as quadrivalent carrying a positive charge or are only trivalent. Descriptions of a few typical searches follow. Searches along carbocyclic peripheries proved of special interest (sections iv and v).

(i) Glycine Residues. Search for



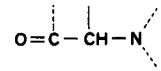
gave eight hits, including the unusual compound **2** (moiety A),



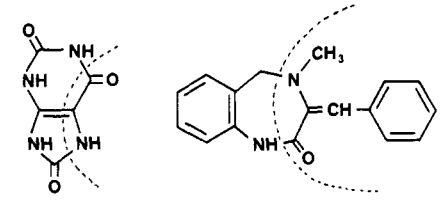
2

which is discussed further under (ii) and (iv) below.

(ii) Amino Acid Residues Having One α-H Atom. Search for



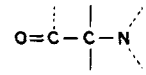
gave 56 hits. This search provided a reminder that such compounds as uric acid (**3**) contain an embedded α-amino acid



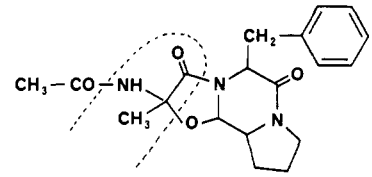
3 **4**

moiety and revealed the benzodiazepinone drug **4** as a derivative of *N*-methylphenylalanine. Compound **2** was likewise a hit, due to moiety B.

(iii) Amino Acid Residues Lacking α-H Atoms. Search for



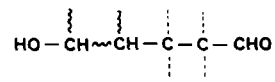
gave eight hits among which, besides some expected 5,5-disubstituted hydantoin drugs, was an ergot peptide alkaloid (**5**)



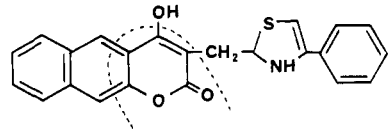
5

containing an α-amino-α-hydroxycarboxylic acid residue, to date uniquely present in this class of compound.

(iv) Coumarins, Dihydrocoumarins, Isocoumarins, etc. Search for

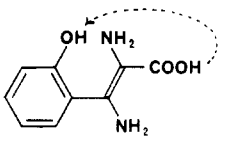


gave 13 hits, of which 12 had the ring structure sought, including the unusual naphthocoumarin **6**. The other hit was



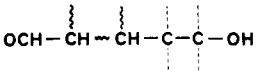
6

compound **2** (above), moieties A and B of which had been found in searches i and ii, respectively. In that light, **2** is a cyclized form of an "opine" amino acid. In the present context, it is a cyclized and hydrogenated derivative of α,β-diamino-*o*-coumaric acid (**7**) from which alternative cyclization, as

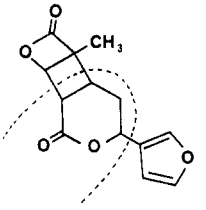


7

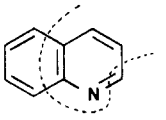
indicated, could yield a diaminocoumarin. A search for isocoumarins using the partial structure



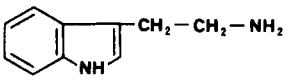
gave only one hit, the unusual compound **8**.



8

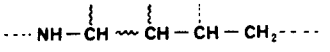


9

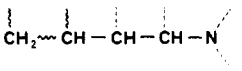


10

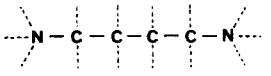
(v) Indole- and Tryptamine-Related Compounds. Search for



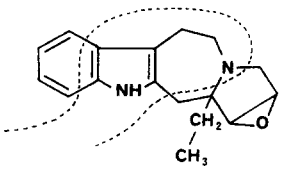
gave 37 hits, of which several were unwanted derivatives of quinoline (**9**). A second pass for



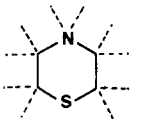
(which would probably have been better done first) reduced these hits to 22 compounds, all containing indolic structures. A good proportion of these had obvious affinities with tryptamine (**10**), and a third pass for



gave, besides **10**, itself and some simple derivatives thereof, several such convoluted compounds as **11**.

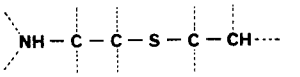


11

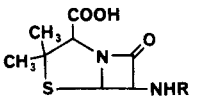


12

(vi) 1,4-Thiazines (without Carbocyclic Fusion). Structure **12** was first searched for, with the string

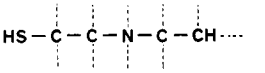


and gave six hits, of which five were penicillins (**13**). The

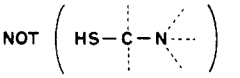


13

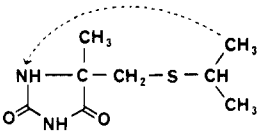
alternative string



would have worked no better. The penicillins were eliminated by a second pass for



and there remained the hydantoin **14**, which, while not con-



14

taining the sought ring **12**, could easily cyclize, as indicated, to yield it.

(vii) Other Searches and Observations. Searches for hydrazines, ureas, thioureas, γ-lactones, thiophenes, thiazoles (carbocyclically fused or not), aldohexosides, ketohexosides, spermine–spermidine-related compounds, glutamic acid derivatives, and lignans have also been carried out, with outcomes similar to those reported above, provided that the class was well enough represented in the search file. Second and third passes with short and common atom strings could lead to "false drops", by finding the sought string in a part of the heterocyclic domain irrelevant to the sought structure. The well-known risk of losing sought structures by "NOT" passes must be kept in mind. It has been particularly interesting, when searches have been aimed at embedded ring structures, quite often to come upon other structures (e.g., **2** and **14**) that, though lacking the sought ring, could well yield it by cyclization or recyclization.

GENERAL CONCLUSIONS

The tree graphs generated by the presented algorithms differ from those generated from cyclic compounds by Randić and Wilkins[3] in that each non H atom of the parent compound is represented once and once only in the tree. This permits comprehensive listing of all paths in the tree,[10] without having to develop the more complicated algorithms needed when the same atom appears more than once in the graph being searched.[3]

Of course, the CAS ONLINE[12] and DARC[13] systems

should achieve the same results as those now described, if the query substructure is properly formulated. However, these systems are not yet in general use for searching privately held files.

The few string searches so far done on the output of the present suite of programs have thrown up a sufficient number of suggestive heterocyclic–aliphatic relationships to justify pursuing this approach more thoroughly. Tree-path tapes (see Figure 7) could be prepared on a large computer from WLN as issued by *Index Chemicus* (or from private files). These could then be searched at leisure on a small computer (with rapid input/output), to follow out ideas as they cropped up, by any chemist familiar with the CROSSBOW bonded-atom symbols and capable of writing very simple string-search programs.

ACKNOWLEDGMENT

I am grateful to Prof. P. M. Stocker and his staff at The Computing Centre of this University (and specially to Dr. P. Anstey and R. A. Jenyon) for much constructive advice about using their facilities, to Prof. M. F. Lynch and his colleagues at The Department of Information Science of Sheffield University (particularly Drs. J. M. Barnard and P. Willett), Pamela A. Chubb, and Dr. Wendy A. Warr for varied and valued advice, and to the Deans and staff of this School for office space and much day-to-day help.

REFERENCES AND NOTES

(1) Robinson, R. "The Structural Relations of Natural Products"; Clarendon Press: Oxford, 1975. Todd, Lord; Cornforth, J. W. *Biogr. Mem. Fellows R. Soc.* **1976**, *22*, 415–527.
(2) Prager, B.; Jacobson, P., Eds. "Beilsteins Handbuch der organischen Chemie," 4th ed.; Springer: Berlin, 1918; Vol 1, p xvi.
(3) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 23–31, 31–37.
(4) Wilson, R. J. "Introduction to Graph Theory"; Longman: London, 1972.
(5) Ash, J. E. In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Horwood: Chichester, England, 1975; Chapter 11, pp 156–176.
(6) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43–48.
(7) The marketing and development rights for the CROSSBOW suite of programs are held by Fraser Williams (Scientific Systems) Ltd., Poynton, Cheshire, England.
(8) Lynch, M. F. *J. Chem. Doc.* **1968**, *8*, 130–133.
(9) Bond, V. B.; Bowman, C. M.; Davison, L. C.; Roush, P. F.; Young, L. F. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 103–105. Warr, W. A. "Proceedings of the CNA (UK) Seminar on Chemical Structure Searching of the Published Literature", March 17–19, 1980, Daresbury, Warrington; Chemical Structure Association: London, 1983; pp 165–80.
(10) Page, E. S.; Wilson, L. B. "Information Representation and Manipulation in a Computer"; Cambridge University Press: Cambridge, England, 1973; p 124, algorithm B.
(11) Windholz, M.; Budavari, S.; Stroumtsos, L. Y.; Fertig, M. N., Eds. "The Merck Index—An Encyclopedia of Chemicals and Drugs", 9th ed.; Merck: Rahway, NJ, 1976.
(12) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(13) Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

Structured Biological Data in the Molecular Access System

SANDOR BARCZA,* LAWRENCE A. KELLY, SIEGFRIED S. WAHRMAN, and RICHARD E. KIRSCHENBAUM

Preclinical Research, Sandoz Research Institute, East Hanover, New Jersey 07936

Received June 11, 1984

Chemical, administrative, and biological information at Sandoz Inc. Research and Development was put into a database created with the MACCS program.[1,2] The configuration of the database and of the "datatypes" in it was done in a way that made the essentially "flat" original design of the database hierarchically structured and searchable. This was accomplished by two devices: (1) The biological activity datatypes were given structured names. The characters went from left (broadest category) to right (most specific category), expressing the major disease goal, then the subgoal, and finally the actual test name. (2) The data within the datatypes were structured into zones and subzones of columns, corresponding to species, dose, effect, direction, date, etc., for each line, while the rows of entry were successive instances of testing. This additional organization of the data offered significant advantages in economy of storage, coherence (interrelatedness) of data, searching, user comprehension, and overview. The orderly entry of data into this system was assured through a data entry interface to the MACCS program. It is the purpose of this paper to describe the innovative adaptation of MACCS to the handling of pharmacologic data, as well as some associated problems and solutions.

INTRODUCTION

Every organization that makes decisions on the basis of data obtained from biologically active substances is confronted with the problem of organizing and retrieving a multiplicity of data elements on a large number of compounds. Sandoz Pharmaceuticals of New Jersey selected MACCS[2] (the Molecular Access System), a data management program based on molecular structures, to manage its large chemical and biological data base. Developed by Molecular Design Limited (MDL), MACCS was chosen because it offered the best commercially available system capable of storing, searching, and retrieving both molecular structure information and associated data.

Sandoz's data management system had to accommodate information derived from over 300 tests on approximately 20 000 compounds—nearly 300 000 lines of information 94 characters wide (27 megabytes). Commitment was made to store chemical and biological information together, in accordance with modern drug research needs. Storage of both chemical and biological information had to be open-ended, allowing Sandoz to add both compounds and data fields as needed. Sandoz developed a chemical and biological information system with MACCS that is graphical, interactive, and

0095-2338/85/1625-0055$01.50/0