10), retrieved by the SSS system and of interest to the searcher, was missed by the RAPID strategy because the carbonyl group was not sought.

In general, a fragmentation code is quite limited in its ability to perform true substructure searches. To locate the co-occurrence of certain atoms in a specific relationship requires a great deal of ingenuity on the part of the RAPID searcher. This capability is built into the SSS system; any substructure that can be adequately defined can be coded and found through SSS.

Another search that illustrates the diversification offered by a topological coding system pertains to compound (XVII) (Figure 11). In looking for related compounds, the RAPID search utilized two approaches; the first strategy sought any compound having the same ring system as compound (XVII) irrespective of substitution. The second strategy looked for those compounds containing a 6,7-dihydroxy- or a 6,7-dimethoxyisoquinoline ring with a diethylamide group two carbon atoms removed from the nitrogen atom of the ring. Although the best available strategies were devised, a negative search resulted. The same substructure was coded for SSS and one of the hits obtained was compound (XVIII) (Figure 12).

SUMMARY AND CONCLUSIONS

Analysis of the results of this comparative study convincingly demonstrated the greater capability of the CAS Substructure Search System over the Frome and O'Day fragmentation code in retrieving chemical compounds. The degree of specificity available in phrasing a query with the SSS system is greater than with this fragmentation code. This stems from the fact that SSS utilizes more fragments (1800 vs. 892) in screening compounds and, in addition, has an iterative search capability for those compounds passing the fragment screen. The topological search enhances the ability of the system to retrieve only pertinent answers. In the present study, 25% of the answers retrieved by RAPID were false drops; the SSS system did not lead to a single false drop.

An over-all comparison of the two methods of chemical searching showed that the Frome and O'Day fragmentation code demands more skill on the part of the searcher, That is to say, the searcher must be very familiar with the compounds in the data base and how they were encoded, he must anticipate every possible variation in functional groups or ring systems, and he must manually eliminate false drops in most searches.

The prejudgment of compounds in the data base that would be of interest in a chemical search was not required with the Substructure Search System. The only requirement was a precise definition of the chemical substructure desired in the compounds being searched.

ACKNOWLEDGMENT

The authors would like to show appreciation for the assistance provided by the personnel in the Statistical Data Branch/Bureau of Drugs: Joyce Hinckley, Carlos Smith, and Henri Williams. Thanks also to Henry Kissman, Alan Gelberg, Bruno Vasta, and Gerard Guthrie of the Science Information Facility for their encouragement and technical guidance. Finally, we acknowledge the valuable cooperative efforts of Charles E. Simmons, Elizabeth McNamee, and Eloise Ingram of the Division of Data Processing.

LITERATURE CITED

(1) Frome, J., and P. T. O'Day, "A General Chemical Compound Code Sheet Format," *J. Chem. Doc.*, **4**, 33–42 (1964).

(2) Gluck, D. J., "A Chemical Structure Storage and Search System Developed at Du Pont," *J. Chem. Doc.*, **5**, 43–51 (1965).

(3) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.*, **5**, 107–13 (1965).

(4) Vasta, B. M., M. L. Spann, and G. T. Guthrie, "Experience with the CAS Substructure System." Paper presented at Fourth Middle Atlantic Regional Meeting, ACS, Washington, D.C., 1969.

Chemical Structure and Substructure Search by Set Reduction*

TAO-KUANG MING and STEPHEN J. TAUBER
National Bureau of Standards, Washington, D. C. 20234

Received April 22, 1970

As part of a computerized system for handling chemical and related information[1] we have included routines for handling chemical structure information. Among these routines is one for structure and substructure search by set reduction, based directly on the method of Sussenguth.[2,3] We have introduced the following refinements: separation of structure search and substructure search into distinct subroutines and inclusion of first-order degree and second-order degree[2] in the control vector for use in structure search. This feature performs much the same function as the connectivity code described by Penny.[4] In addition our routine has processed structures with greater symmetry than any which Sussenguth's routine is

* Presented in part before the Computers Division of the 3rd Great Lakes Regional Meeting, ACS, DeKalb, Ill., June 5–6, 1969.

This work was supported by the National Institutes of Health under an interagency agreement. Contribution from the National Bureau of Standards. Not subject to copyright.

The set reduction method by Sussenguth for chemical structure and substructure searching has been adapted: Structure search and substructure search have been separated into distinct subroutines; first-order and second-order degree have been included in the control vector. This method has been tested on structures with greater symmetry than has been done previously, and it has been verified for disjoint substructures. A method has been demonstrated for removing the ambiguity inherent in using a bond adjacency matrix.

reported to have been tested on. We have also demonstrated that the routine can be used for seeking the co-occurrence of two disjoint substructures.

The complete structure search is of course the (full) graph matching algorithm. For the substructure search we have used the partial subgraph matching algorithm, rather than that for subgraph matching; the former algorithm accepts both partial graphs and partial subgraphs. We thus accept envelopes of fused-ring systems as substructures and similarly all the atoms of a ring without necessarily all of the bonds (cf. Table I). The node properties[3] used for defining the sets are atomic number, bond value (single, double, triple, or "R-type"[5]), first-order degree, and second-order degree. When a set containing only one node appears, the sets of adjacent nodes and of nodes one node removed are also introduced. Hydrogen atoms attached to carbon are omitted.

The program was written in Fortran V[6] and runs on the Univac 1108. It consists of about 2250 source language statements, with DO loops nested as much as 6 layers deep. Together with the necessary operating system subroutines it occupies about 31,400 words of core memory (21,850 for instructions, 9550 for data).

SET CORRESPONDENCE

The reason for separating the structure search and the substructure search is chiefly the difference between the ways in which the complementing of sets operates in the two types of search. Let the query have n nodes $(x_1, x_2, \ldots, x_n)$; let the query set Q_i be defined by property p_i and contain j members.

$$Q_i = \{x_{i_1}, x_{i_2}, \ldots, x_{i_j}\} \qquad j \leq n;$$

let the disclosure have m nodes $(y_1, y_2, \ldots, y_m)$; and let the disclosure set D_i be defined by the corresponding property p_i' and contain k members

$$D_i = \{y_{i_1}, y_{i_2}, \ldots, y_{i_k}\} \qquad k \leq m.$$

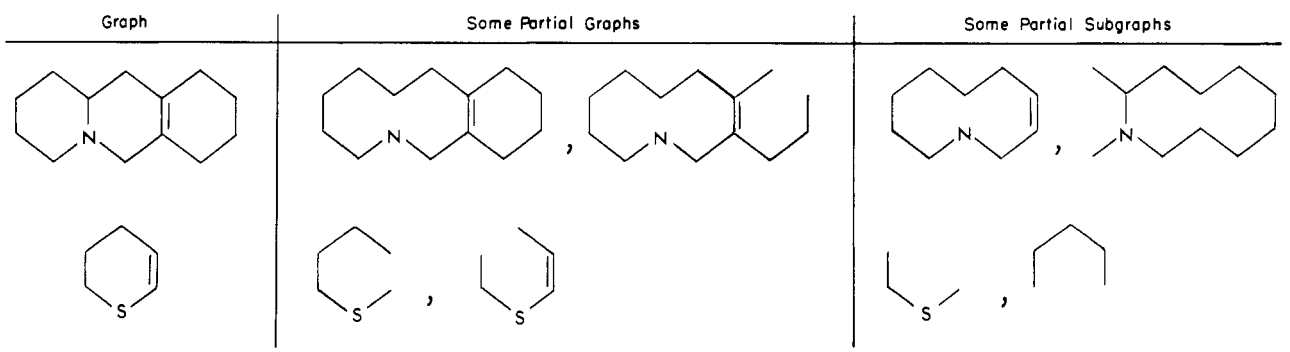
For some properties p_i the corresponding properties p_i' are identical, e.g., atomic number. However, in the partial subgraph matching algorithm some of the properties of the disclosure nodes differ from the properties of the query nodes with which they are put in correspondence. For example, the first-order and the second-order degrees may validly be greater for the disclosure nodes. The algorithm establishes as corresponding to each other the set of query nodes with degree equal to l and the set of disclosure nodes with degree equal to or greater than l. Consequently there can exist corresponding sets Q_i and D_i for which the complements $\bar{Q}_i$ and $\bar{D}_i$ do not correspond. The correspondence (symbolized by $\leftrightarrow$) between $\bar{Q}_i$ and $\bar{D}_i$ can be assumed only if the number of members of the original sets are equal:

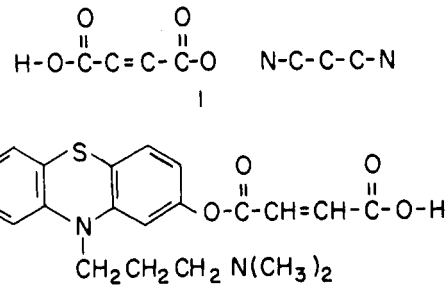
$$[N(Q_i) = N(D_i)] \Rightarrow (\bar{Q}_i \leftrightarrow \bar{D}_i).$$

This condition is always met in graph matching when a mismatch has not yet been detected. By contrast it must be tested for in partial subgraph matching.[2] Since this test occurs in one of the innermost repetitive loops of the program its elimination significantly increases the efficiency of structure matching. There are also differences in the assignment procedure[3] between graph matching and partial subgraph matching because in a successful partial subgraph match some disclosure nodes may remain unused. The routine can terminate after matching $n - 1$ nodes, instead of all n, only in a full structure search.

Sussenguth barely alludes to disjoint subgraphs, in relation to Cayley Colour Graphs,[2] and he does not exploit the power of his algorithm to search simultaneously for two co-occurring chemical substructures. The definitions of the sets used in the partial subgraph matching algorithm does not depend on the connectedness of the partial subgraph (nor of the graph). Consequently a disjoint graph, representing two or more distinct chemical substructures, can be posed as a query. Thus, the pair of substructures I is found in structure II. It is important to recognize that an atom cannot serve double duty; the substructures being sought must coexist without overlap-

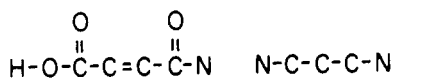
Table I.



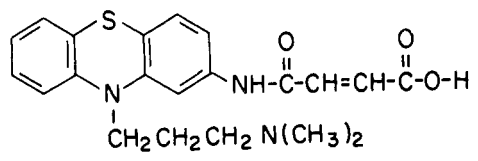


I

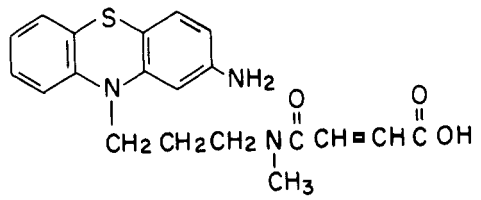
II



III



IV



V

ping. Thus substructure pair III is found in structure IV but not in V.

ASSIGNMENT

When set operations can effect no further reduction, but there nevertheless are sets with more than one member, then the assignment procedure is invoked.[3] This may either help to distinguish nodes which differ with respect to properties which are ignored by our implementation of the algorithm or arbitrarily assign to a matching pair one node of a set of entirely identical nodes.

Sussenguth's tests were limited to those chemical structures which have at most binary symmetry.[2] Thus in structures VI and VII the following sets result:

$Q_1 = \{c, d\}$	$D_1 = \{1, 4\}$
$Q_2 = \{a, b, e, f\}$	$D_2 = \{2, 3, 5, 6\}$.

One assignment completely resolves sets Q_1 and D_1, at the same time partially resolving sets Q_2 and D_2 into subsets which can each be completely resolved into subsets by one further assignment:

Arbitrary assignment:	$Q_3 = \{c\}$	$D_3 = \{1\}$
Consequent resolutions:	$Q_4 = \{d\}$	$D_4 = \{4\}$
	$Q_5 = \{a, b\}$	$D_5 = \{2, 3\}$
	$Q_6 = \{e, f\}$	$D_6 = \{5, 6\}$

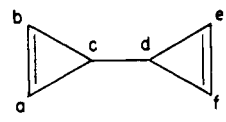
By contrast, in structures VIII and IX several assignments are needed to resolve the following sets Q_2, Q_3, D_2, and D_3:

$Q_1 = \{a\}$		$D_1 = \{3\}$
$Q_2 = \{b, f, j\}$		$D_2 = \{2, 4, 8, 12\}$
$Q_3 = \{c, d, e, g, h, i, k, l, m\}$		$D_3 = \{1, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17\}$
Arbitrary assignment:	$Q_4 = \{b\}$	$D_4 = \{2\}$
Resultant sets:	$Q_5 = \{f, j\}$	$D_5 = \{4, 8, 12\}$
	$Q_6 = \{c, d, e\}$	$D_6 = \{1, 6, 7\}$
	$Q_7 = \{g, h, i, k, l, m\}$	$D_7 = \{5, 9, 10, 11, 13, 14, 15, 16, 17\}$

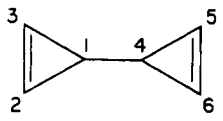
Even when sets Q_5 and D_5 have been completely resolved by two arbitrary assignments, sets Q_6 and D_6 plus the two other pairs of subsets of sets Q_3 and D_3 must each be resolved by arbitrary assignments.

Whenever a cascade of arbitrary assignments can occur (such as is necessitated by symmetrically situated *tert*-butyl groups), it is also necessary to provide back-up procedures to recover from erroneous optional assignments which have been introduced not by the symmetry of the structure in question but by the use of insufficiently powerful node properties. For this purpose we have used a push-down stack which permits recovery from arbitrary assignments which lead to contradictions in the inverse of the order in which the assignments were made. For example, in attempting to map substructure X into structure XI more than one arbitrary assignment is necessary:

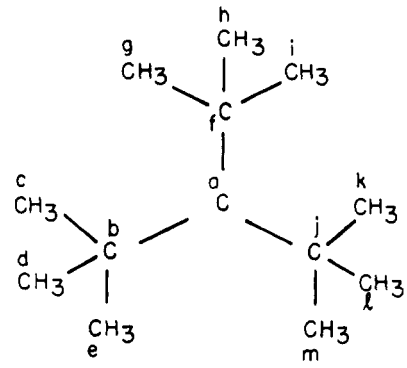
Assignment 1:	$Q_1 = \{a\}$	$D_1 = \{1\}$
Consequent resolution:	$Q_2 = \{b, c\}$	$D_2 = \{2, 4, 5\}$
Assignment 2:	$Q_2 = \{b\}$	$D_3 = \{2\}$
Consequent resolution:	$Q_4 = \{c\}$	$D_4 = \Lambda$



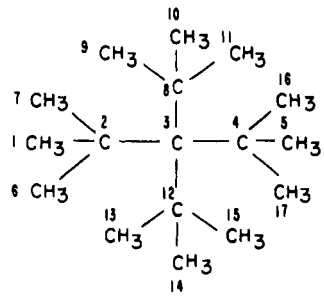
VI



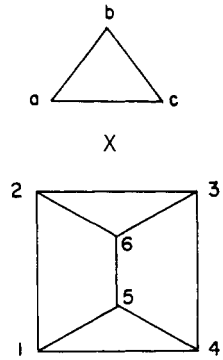
VII



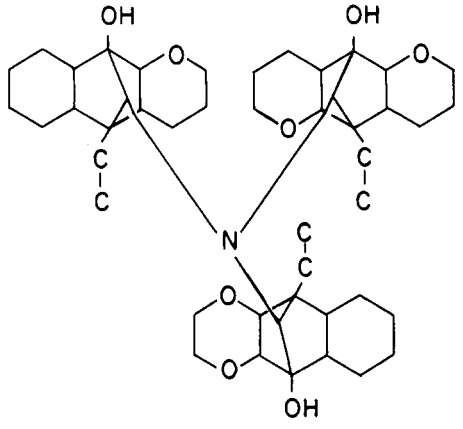
VIII



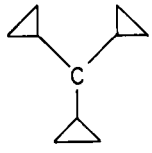
IX



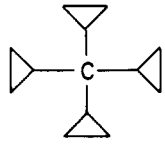
X XI



XII



XIII



XIV

XV

Assignment 2 results in a contradiction, since set D_4 is empty. Consequently assignment 2 must be rescinded but assignment 1 is to be kept.

Assignment 1:	$Q_1 = \{a\}$	$D_1 = \{1\}$
Assignment 2′:	$Q_5 = \{b\}$	$D_5 = \{4\}$
Consequent resolution:	$Q_6 = \{c\}$	$D_6 = \{5\}$

Only if all alternative assignments at a lower level fail is the assignment at the higher level rescinded. The push-down stack stores pointers to the locations in the working area where the arbitrarily assigned sets occur.

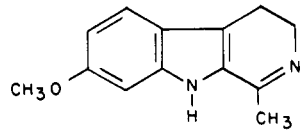
It is important to recognize that when many assignments are made this algorithm degenerates into a form of iterative search, tentatively pairing various nodes in query and disclosure sets. The consequences of each arbitrary assignment are propagated by the algorithm until the algorithm terminates or until a further assignment is needed. Thus, the time required for searches can rapidly increase with the level of symmetry of the structures, and it is not clear that our algorithm is economically feasible.

TIMING

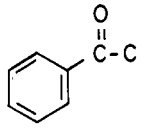
Our results agree entirely with those of Sussenguth[2] in the following two respects: A complete match for a large molecule takes considerably longer than for a small one and a mismatch is generally found more quickly than a match.

A complete match of one representation of structure XI (6 atoms) against another required 0.58 sec; Sussenguth's structure 2602[6] without hydrogens (28 atoms) required 148 sec; structure XII (58 atoms) required 196 sec. Since the higher degrees of symmetry can exist only in large structures, careful study of the effect of the extent of symmetry would require many searches with large query and disclosure molecules. These searches, requiring 2 and 3 minutes, were sufficiently costly to dissuade us from investing in such a study. We did determine that structure XIII as a query was found as a substructure of XIV in 7.4 sec. Structure XV was found in XI in 5.2 sec, including set-up time; mismatch against structure XVI took 0.82 sec.

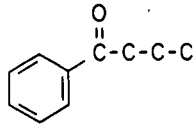
Several substructure searches have been run on a file of psychotropic drugs[9] with the following results: Query structure XVII required from 14 to 118 sec to find matches against disclosures of 35 to 41 atoms; it found a mismatch against a disclosure of 32 atoms in 0.20 sec, and it found 4 mismatches on control vector alone in 0.28 sec. Query structure pair I found matches in 32- to 41-atom disclosures in 6.3 to 41 sec and a mismatch against a 40-atom structure in 0.23 sec. Query structure XVIII found in 26 sec that there was no match in a 60-compound portion of the file. Finally a set of complex queries[1] was put to the file in such a way as to cause 4 substructure searches during a single computer run: Query structures XIX, XX, XXI, and XXII were tested against 5, 5, 145, and 8 disclosures respectively in a total of 21 sec.



XVI



XVII



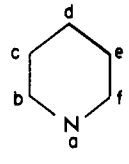
XVIII

C-O-C=O XIX

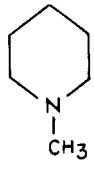
C-O-P XX

C-C-N-C XXI

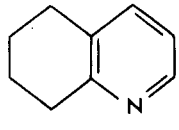
O-P-O XXII



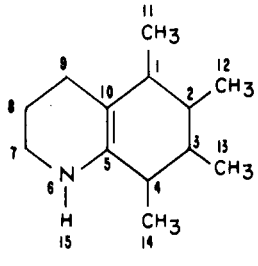
XXIII



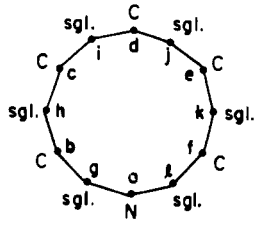
XXIV



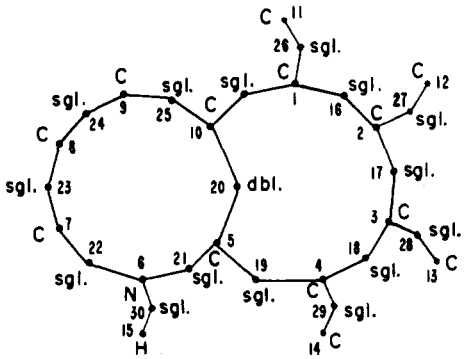
XXV



XXVI



XXVII



XXVIII

INHERENT AMBIGUITY

The one inherent weakness noted by Jochelson et al.[7] in the Sussenguth set reduction algorithm remains in our implementation: Bonds do not appear in their own right; bond values appear instead in a bond adjacency matrix[3] which states the values of all bonds adjacent to each atom. False drops can therefore occur, particularly in a substructure search. In searching for substructure XXIII in structures XXIV, XXV and XXVI, our program correctly accepts XXIV and correctly rejects XXV (because there are only R-type bonds in the nitrogen-containing ring), but it incorrectly accepts XXVI. This occurs because the algorithm asks, with respect to atoms 5 and 10, only whether each of them is adjacent to a single bond but not whether the bond between them is a single bond. Removing this source of error does not require separate connection matrices for each bond value as suggested by Jochelson et al. It can also be eliminated by explicitly introducing the bonds as nodes of the graph representing the chemical structure. Substructure XXIII and structure XXVI would now be represented as XXVII and XXVIII. The single bond between atoms *e* and *f* is now explicitly cited as node *k*, and the double bond between atoms 5 and 10 is explicitly cited as node 20.

The validity of this approach was verified by usurping atomic numbers 1 and 2 for representing single and double bonds (in structures without hydrogen or helium). If explicit citation of bonds were to be used routinely in this fashion, then the bond adjacency matrix should be eliminated from the implementation; all branches of the graph expanded in this fashion would necessarily have identical values, and the bond adjacency matrix would be devoid of information. The expanded graphs would require approximately double the space for storage. Furthermore the strategy of applying the basic algorithm effectively would be different, because atoms connected to each other are no longer represented by adjacent nodes in the graph; such a strategy would require a considerably different implementation. We decided that during normal use of our over-all system, the proportion of false drops to be expected due to the bond adjacency ambiguity was insufficient to warrant the effort of its elimination.

CONCLUSION

Testing of the over-all system in which the program here described is embedded is continuing. It is evident at this time that the set reduction search is practical for routine use at least with reasonably small substructure queries.

LITERATURE CITED

(1) Marron, B. A., and S. J. Tauber, "Evolution of a General Computer-Based Information System from Pharmacological Requirements," *Proc. Amer. Soc. Info. Sci.*, **6**, 223 (1969).
(2) Sussenguth, E. H., "Structure Matching in Information Processing" (thesis, Harvard Univ., 1964).
(3) Sussenguth, E. H., "A Graph-Matching Algorithm for Matching Chemical Structures," *J. Chem. Doc.*, **6**, 36 (1965).
(4) Penny, R. H., "A Connectivity Code for Use in Describing Chemical Structures," *J. Chem. Doc.*, **5**, 113 (1965).
(5) The precise definition of R rings is part of the Hayward notation rules. As a first approximation, an R ring may be taken as an aromatic ring or one with alternating single and double bonds. There are however R rings that are not aromatic and vice versa. Every bond in an R ring is an R-type bond.
(6) "Univac Data Processing Division 1108 Multi-Processor System FORTRAN V Programmer's Reference Manual," Pub. **UP-4060**, Sperry Rand Corp., 1966.
(7) Jochelson, N., C. M. Mohr, and R. C. Reid, "The Automation of Structural Group Contribution Methods in the Estimation of Physical Properties," *J. Chem. Doc.*, **8**, 113 (1968).
(8) Ref. 2, pp 5–61.
(9) Usdin, E., and D. H. Efron, "Psychotropic Drugs and Related Compounds," Public Health Service Pub. **1589**, Washington, D. C., 1967.