ARTICLES

The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens

P. G. DITTMAR, N. A. FARMER,* W. FISANICK, R. C. HAINES, and J. MOCKUS
Chemical Abstracts Service, Columbus, Ohio 43210

Received October 8, 1982

The CAS ONLINE search system provides on-line access to the more than 6 million substances in the Chemical Abstracts Service (CAS) Chemical Registry System files. Both substructure and full-structure searches can be performed. Search queries are normally input as structure diagrams. An initial screen search checks each file substance for the presence of specified structural features (screens) and eliminates those substances not containing all of the requisite structural features; an atom-by-atom search then checks each candidate answer against the query definition, accepting as answers only those substances that match exactly. Since atom-by-atom searching is a relatively slow procedure, efficient screen searching is necessary to minimize the number of answers passed on to the second search step. Efficiency is achieved by careful selection of the structural features used as screens. This paper discusses the history of CAS search systems, the basic system design and the "search machine" approach, techniques for screen searching, the types of screens used by CAS ONLINE, the screen dictionary (the authority list of screens), screen generation for the CAS ONLINE search files, and the encoding of screen search queries.

INTRODUCTION

The CAS ONLINE Search System provides on-line access to the chemical structure and nomenclature data bases of the Chemical Abstracts Service (CAS) Chemical Registry System.[1,2] Full structure searches (searches for precise structure matches or for families of closely related structures) and substructure searches can be conducted on line through remote terminals, searching the file of more than 6 million substances that have been indexed in *Chemical Abstracts* (CA) since 1965. Each search answer provides a structure diagram, a CAS Registry Number, a CA index name and up to 50 synonyms (including common and trade names), and, at the searcher's option, CA abstract numbers and bibliographic references for the ten most recent references to the substance in the literature. During 1983, CAS will add the capabilities to display abstracts for some 3.5 million documents and to search CAS bibliographic and index data back to 1967.

In the initial version of CAS ONLINE, introduced late in 1980, substructure search questions were framed in terms of screen numbers selected from a dictionary that identified almost 6000 different structural features. Screens could be combined by using Boolean logic operators to represent almost any desired structure or substructure, and the CAS ONLINE system would retrieve as an answer any substance described by the specified combination of screens.

In late 1981, the CAS ONLINE system was upgraded to simplify and improve the query-framing and search procedures. The searcher no longer has to analyze a search query in terms of search screens, since a new query framing procedure allows a query to be defined in terms of a structure diagram, constructed on either a graphics or a text terminal. The system itself will identify the screens to be used in the screen search and then perform the search. Finally, a new atom-by-atom search procedure checks each structure retrieved by the screen search and passes as answers only those which precisely match the search query. This matching procedure eliminates the major drawback of the initial system, the retrieval of "false drops" containing the specified structural fragments in other than the desired relationship.

Although the CAS ONLINE searcher no longer needs to frame the typical search query in terms of screens (nor for the most part even needs to know that screens are used by the system), search screens still are and will remain a fundamental part of the CAS ONLINE system. This paper describes the basic design of the system, and discusses screen search techniques, the types of screens used by CAS ONLINE, and the selection of the screens used to characterize each substance on file. Later papers will discuss query structure input, the command language and user interface, and search capabilities.

HISTORY OF CAS SEARCH SYSTEMS

CAS involvement with substructure search systems began in the late 1960s, with the development of an experimental system to search Registry II structure files.[3] This system had screen and atom-by-atom search capabilities comparable to those of the present CAS ONLINE system but was restricted by the computer technology of the time; it was a tape-based batch mode system, with screen and atom-by-atom search queries manually encoded by the searcher and with answers limited to Registry Numbers. While the system did show the feasibility of substructure search procedures built around the CAS Registry structure files, it was hampered by lack of sufficiently powerful computer hardware and by the inability to provide structure diagrams for answers. As a consequence, the system was never promoted as a CAS service, although it was released on an experimental basis to several outside organizations. It was also used by the National Cancer Institute in a chemical information system to support their biological screening programs.[4]

A few years later, several Swiss chemical firms—Ciba Ltd., J. R. Geigy Ltd. (now Ciba-Geigy Ltd.), F. Hoffmann-La Roche & Co. Ltd., and Sandoz Ltd.—decided to jointly develop a computer system for chemical information based on the CAS Registry system and formed the Basel Information Center for Chemistry (BASIC).[5] Their system used the experimental CAS substructure search system to search a CAS Registry structure file licensed from CAS as well as private structure files built and maintained by a system based on CAS

0095-2338/83/1623-0093$01.50/0

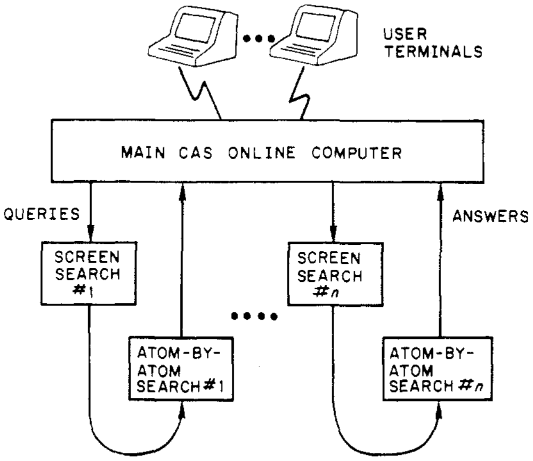


Figure 1. Architecture of the search machine.

Registry II programs. The BASIC group made a number of improvements to the original CAS search system, developing several new screen types and revising the screen dictionary; these improvements reduced the costs of search screen generation and (by providing better screen-out) atom-by-atom searching, the two time-consuming aspects of substructure search.[6,7] At the same time, BASIC upgraded the private structure file aspects of their system with the CAS Registry III programs.[6]

In the late 1970s, CAS reexamined the possibilities of substructure search as a CAS service. While the earlier work discussed above had demonstrated the feasibility of using an initial screen search followed by atom-by-atom search, it was clear that the large size of the file (then already over 5 million substances) would present problems. The inverted file organization usually used for on-line searching could be expected to lead to unacceptably slow response times, due to the very large file size. Some of the screens which would normally be generated from a query would be assigned to very large numbers of substances, leading to the necessity to intersect very large lists of potential retrievals.

The result of the investigation was a new approach to large-file searching.[8,9] Under this approach, the file is segmented and the search task distributed over a number of minicomputers operating in parallel, the "search machine". Each minicomputer searches only a portion of the file, using a simple sequential file organization. A "front end" controller handles all of the system overhead tasks; query input, supervision of the search machine, and answer output. A simplified view of this architecture is shown in Figure 1. (This approach has subsequently been adapted to an industrial substance search system by Hagadone and Howe[10].)

The segmentation of the search files provides the potential for a very high speed search. The screen search procedures are designed to overlap processing and data input: while one block of data is being searched, the next is being read into the minicomputer. By careful "tuning" of the search program and the search file organization, the screen search can operate as fast as data can be read from the magnetic disks providing file storage. As a result, the time required to perform a screen search does not depend upon the size of the full file but upon the size of a file segment searched by a single minicomputer. This means that any given response time for the initial screen search can be achieved simply by choosing the appropriate size for file segments.

The search functions are also divided so that the screen search and atom-by-atom search are performed by two connected sets of minicomputers. Atom-by-atom searching is a relatively slow process compared to the Boolean logic operations of screen searching, since it must match (by an iterative path-tracing procedure) the query definition against the connection table structure record of each potential answer passed by the screen search. By pooling the answers provided by the screen search minicomputers and distributing them to a second set of minicomputers performing the atom-by-atom search, the computer resources can be adjusted to provide an acceptable response time for the typical substructure search. If response time is judged to be too slow, the process can be speeded simply by adding another minicomputer to do atom-by-atom searching; there is no reason to have the same number of minicomputers performing each function.

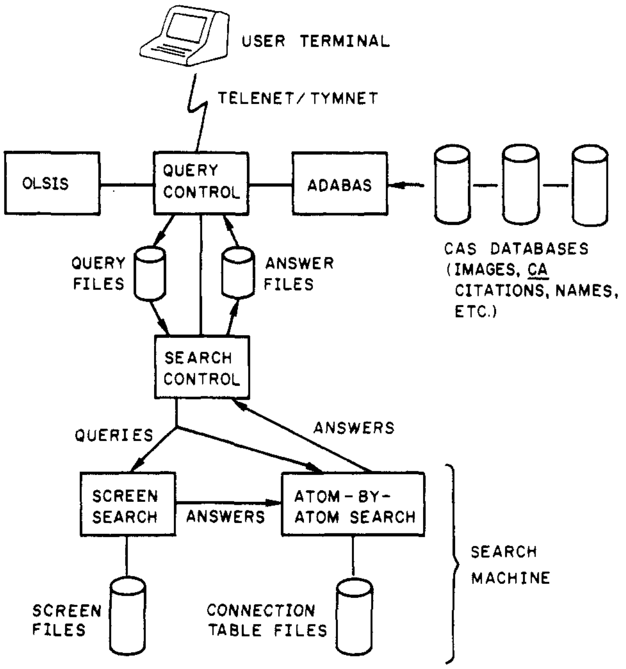


Figure 2. Organization of the CAS ONLINE search system.

The new search machine architecture provides three additional benefits, besides making search response times independent of file size. First, the "pipeline" flow, where answers continually move through the system, allows answers to be displayed for review by the user as they are found; there is no need to wait several minutes to see the results of a search. Second, the sequential organization of all of the files allows very easy updating, so that search files can be updated weekly to provide the CAS ONLINE searcher with almost immediate access to the approximately 7000 new substances added to the CAS data bases each week. Finally, multiple users can be handled simultaneously without significant degradation of system response times.

GENERAL SYSTEM DESIGN

The general organization of the CAS ONLINE search system is shown in Figure 2.

The search machine computer hardware facilities of the present CAS ONLINE system consist of a set of Digital Equipment Corp. PDP 11/45 minicomputers for screen searches, a set of PDP 11/44 minicomputers for atom-by-atom searches, and a PDP 11/55 and a PDP 11/44 minicomputer for system support, for maintenance, and, most importantly, for backup of the search minicomputers. Each search minicomputer has its own 300 Mbyte disk holding search data; currently, each minicomputer has data for approximately 750 000 substances.

The CAS ONLINE system also uses some of the resources of a large computer system, currently an IBM 3081. The CAS ONLINE programs operating on this system include the telecommunications interface, the query and search control procedures, a version of the Online Structure Input System

(OLSIS) used for graphic structure input,[11] the ADABAS data-base management program (a product of Software AG) used during answer data retrieval to access the CAS production data bases, and an output program for off-line prints through a high-speed Xerox 9700 laser printer.

Other programs also provide behind-the-scenes support for the CAS ONLINE system. These include the fragment generation and search screen generation programs used to build the search files for the screen search, and the Algorithmic Structure Display program used to create the image file of structure diagrams for search answers.[12]

SCREEN SEARCH TECHNIQUES

The CAS ONLINE system allows a search query to be defined in terms of one or several structure diagrams and/or manually encoded screen sets, with components combined by the AND, OR, and NOT Boolean operators. Query structure input, which is the usual way in which queries are presented to the system, provides fast and easy definition of a search query and also allows an additional atom-by-atom matching of answers against the query, resulting in highly precise searches. However, there are circumstances in which the use of manually encoded screen sets to describe either an entire query or part of a query is desirable. Whether a screen search query is manually encoded or automatically generated, it is defined at the lowest level in terms of screen sets: single screens or simple Boolean expressions (i.e., without parentheses) of screens connected by AND and OR operators. These screen sets are then grouped in a Boolean expression to form a query statement in which the AND, OR, and NOT operators may be used, and complex expressions may be formed with parentheses.

The screen search procedures of the present CAS ONLINE system use 2128-bit strings to record screen information for searching. There is a direct one-to-one correspondence between bits and screen numbers, although many screen numbers are associated with several screen definitions. This screen number sharing, discussed in more detail below, was done to increase the number of search screens available to the searcher; almost 6000 different screen definitions are present in the CAS ONLINE screen dictionary.

SCREENS FOR SCREEN SEARCHING

The selection of screens for a substructure search system has long been a subject for research, and many studies have been done.[6,7,13,14] The design of a "best" screen set, though, is very highly dependent upon the nature of the structures being searched and the types of queries. Further, the need for a "best" screen set is reduced by the addition of a second, atom-by-atom, search step to precisely match potential answers against the search query. Once this is done, the role of the screen search changes from answer retrieval to elimination of substances which are not answers, so that the atom-by-atom search will operate efficiently.

Given these considerations and the desired performance criteria that had been established for the search machine, CAS chose to use a screen set based upon one developed by BASIC, which had, in turn, been developed from a screen set used in early CAS investigations of substructure search.[6,7,14] [Although the initial CAS screen set had been developed on an empirical basis (comprehensive coverage of chemically likely structures) rather than extensive statistical analysis,[13,14] it had nevertheless performed well both for CAS and for BASIC.] CAS added a number of generic screens to support less specific search queries defined through structure input. (The availability of these screens simplified the design of the automatic screen generation procedure, since otherwise a generic structural feature would have had to be defined by several specific screens grouped with OR logic.) The addition of these generic screens did not affect the response times of the CAS ONLINE screen search, due to the search machine architecture; however, in a system using an inverted file structure their use could have caused serious problems.

Although the present CAS ONLINE screen set operates at a very acceptable level of performance, CAS has continued to investigate possible additions to the screen set. Additional definitions of existing types of screens as well as new screen types are being considered. The addition of new screens to the CAS ONLINE screen dictionary (the authority list of screens) would only be done if a substantial increase in performance would be provided. Since the addition of new screens would require processing the full 6-million-plus CAS Registry structure file, the benefits provided by any new screens would have to justify the cost of their creation.

One type of screen currently being considered for use in CAS ONLINE is a type of ring screen similar to one used in an early CAS search system based on CA index nomenclature.[15] This screen type would indicate the presence of ring features in a substance and would be derived from the elemental compositions of the ring systems. Some of the screens would indicate the presence of specific ring systems from among the approximately 200 that account for almost 95% of all ring system occurrences in the substances of the CAS Registry system.[16] Other screens would indicate such features as the individual component rings of larger ring systems and generic ring information such as the presence of a carbocyclic ring system.

STRUCTURE FRAGMENTS AND SEARCH SCREENS

In the discussions that follow, the term "fragment" refers to a structural feature of a substance and "screen" to a search term which is assigned a reference number and listed in the CAS ONLINE screen dictionary. (From another viewpoint, a screen is a fragment which has been selected for inclusion in the dictionary because of its usefulness for searching.) This distinction is necessary because some screens indicate the presence of any of a number of different fragments, due to screen number sharing. (This technique is discussed later.)

Most structure fragments are derived from what is termed the "graph" of the connection table, which describes the nonhydrogen atoms and the bonds between them that comprise the basic structure of the substance.[2] Simple derivatives of this basic structure (i.e., hydrates and simple metal or acid salts) that are described via the single atom fragment portion of the Registry record lead to the generation of graph modifier element (GM E) fragments. Other graph modifier fragments are derived from other data elements in the Registry records and indicate the presence of special structural features (e.g., charge, unusual mass), multicomponent substances, etc.

The CAS ONLINE search system screen dictionary is the authority list of screens used by the system. It contains twelve types of screens in three broad classes: (1) augmented atom screens, which describe atoms and their immediate attachments; these include augmented atom (AA), hydrogen augmented atom (HA), and twin augmented atom (TW) screens; (2) linear sequence screens, which describe linear strings of nonhydrogen atoms, e.g., atom sequence (AS), bond sequence (BS), and connectivity sequence (CS) screens; (3) general structural feature screens, e.g., ring count (RC), type of ring (TR), atom count (AC), degree of connectivity (DC), element composition (EC), and graph modifier (GM) screens. (The distribution of screens by type is detailed in Table I.) These screen types are discussed below and illustrated via examples from the structure of 6-chloro-4-(4-hydroxyphenyl)-2-pyridinecarboxylic acid hydrochloride (Figure 3).

Each screen entry in the dictionary includes a count or

Table I. Distribution of Fragments and Screens by Type

fragment/screen type	fragments		screen no.	
AA-augmented atom				
bond type and value	600		314	
bond type only	515		275	
unspecified bond	427		179	
unspecified atom	15		15	
AA-total	1557	1557	783	783
HA-hydrogen augmented atom		296		113
TW-twin augmented atom		30		17
		1883		913
AS-atom sequence				
bond type only	2024		425	
unspecified bond	2153		103	
AS-total	4177	4177	528	528
BS-bond sequence				
bond type and value	921		190	
bond type only	17		17	
BS-total	938	938	207	207
CS-connectivity sequence				
bond type only	364		153	
unspecified bond	87		62	
CS-total	451	451	215	215
		5566		950
AC-atom count		19		19
EC & GME-element count		639		122
DC-degree of connectivity		17		17
RC-ring count		10		10
TR-type of ring		70		51
GM-graph modifier special		37		37
		792		256
augmented atoms total		1883		913
sequences total		5566		950
general structural features total		792		256
total		8241		2119

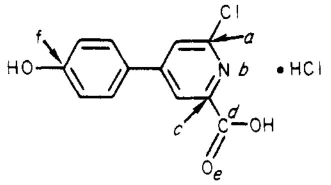


Figure 3. Example structure.

number that specifies the number of occurrences of the associated fragment (or any of a set of fragments) in the structure; a count of one, however, is not explicitly shown. This count is a minimum value, not an exact value, so that, for example, a count of two means that the fragment appears two or more times. Thus, a structure that contains five oxygen atoms would be retrieved by a search query that specifies that three or more oxygen atoms must be present.

Augmented Atoms (AA). These screens are descriptions of atoms and their nonhydrogen attachments. In the description of an AA fragment, the central atom is cited first, followed by its attachments in element symbol order. If bonds are specified, using the bond symbols shown in Table II, a secondary ordering on bond type cites ring bonds (indicated by the asterisk symbol) before chain bonds (indicated by the minus symbol); complete bond specifications of both type and value are cited in the order

*1 *2 *3 *4 -1 -2 -3 -4

Thus, in the example structure, carbon atom c could be described by a number of AA fragments including

```
AA C    C    C    N
AA C *  C -  C *  N
AA C *4 C -1 C *4 N
```

Table II. Bond Symbols

symbol	definition
*	any ring bond (value not defined)
–	any chain bond (value not defined)
*1	single ring bond
–1	single chain bond
*2	double ring bond
–2	double chain bond
*3	triple ring bond
–3	triple chain bond
*4	alternating (aromatic or completely conjugated) or tautomeric or delocalized ring bond[17]
–4	tautomeric or delocalized chain bond[17]

Note that an augmented atom fragment may be a complete description of a central atom and its environment, as in

AA C *1 C *1 C -2 O

or just a partial description, as in

```
AA C -1 C -2 O
AA C *  C *  C - O
```

In the latter case, additional attachments to the central atom may or may not be present. If it is desired to prohibit additional attachments when encoding a query profile, an HA or TW fragment must be used to completely describe the central atom.

This "inclusive" approach, like the "or more" approach to the specification of screen counts, facilitates substructure searching by automatically providing for the retrieval of larger structures that contain the desired substructure. Some other substructure search systems use an "exact" or "exclusive" approach, so that the searcher must explicitly specify all possible alternatives in order to do a truly generic search. While that approach does facilitate highly specific substructure searches and exact-match full-structure searches, it severely hampers more general searches and may, on occasion, lead to incomplete recall.

Hydrogen Augmented Atoms (HA). These screens are augmented atoms whose definitions include a specification of the hydrogen attached to the central atom. This is an exact count of the number of attached hydrogens (including D or T, if present), not an "or more" count. The count follows the central atom's element symbol and appears as "C " for a carbon atom with no attached hydrogens, "C H" for a carbon atom with one hydrogen, "C H2" for a carbon with two hydrogens, etc.

For the example structure, the fragment used to describe the hydroxy group attached to atom f would be

HA O H -1 C

This fragment would not be used to describe the hydroxy group attached to atom d, because HA (and TW) fragments cannot be used to specify the presence of hydrogen on nitrogen or chalcogen (O, S, Se, or Te) atoms involved in tautomeric situations.[17]

Twin Augmented Atoms (TW). These screens are augmented atoms whose definitions include the specification of the hydrogen attached to the central atom and to one of its attached atoms.

For the example structure, the fragment used to describe atom f would be

TW C *4 C *4 C -1 O H

Atom Sequences (AS). These screens are descriptions of linear sequences of four, five, or six nonhydrogen atoms. Bond types may be specified for the more common AS fragments, but bond values are not used.

For the example structure, the a–b–c–d–e atom sequence would be described by either of two equivalent AS fragments:

AS C * N * C - C - O

AS O - C - C * N * C

(As an aid to the searcher, both the "forward" and "reverse" descriptions of such unsymmetrical sequences are provided in the CAS ONLINE screen dictionary.)

Bond Sequences (BS). These screens are descriptions of linear sequences of three, four, or five bonds, always specifying the bond types and often the bond values. In contrast to AS screens, BS screens do not specify the element types of the atoms involved and simply include "A" dummy atom symbols, for clarity, between the bond symbols.

For the example structure, the a–b–c–d–e atom sequence would be described by either of two equivalent BS fragments:

BS A *4 A *4 A -1 A -4 A

BS A -4 A -1 A *4 A *4 A

Connectivity Sequences (CS). These screens are descriptions of nonhydrogen connectivities for linear sequences of four, five, or six atoms, often including bond types but never bond values. The connectivity values here are the exact number of nonhydrogen attachments, not the usual "or more" specification, so that, for example, a "1" would always indicate a terminal atom and a "2" an atom with exactly two nonhydrogen attachments. As a result of this approach, CS screens are of more use for full structure searches than substructure searches, though several related CS screens, grouped with OR logic, may be used for the latter.

For the example structure, the a–b–c–d–e atom sequence would be described by either of two equivalent CS fragments:

CS 3 * 2 * 3 - 3 - 1

CS 1 - 3 - 3 * 2 * 3

Ring Count (RC). These screens specify the minimum number of rings present in the structure, defined as a count of the ring closure pairs present in the structure record and equal to the minimum number of bonds that would have to be broken to open all rings.

The RC 2 screen would be used to describe the example structure.

Type of Ring (TR). These screens describe the node sequences of rings of three to seven atoms. The symbol "D" is used to indicate a nonfused ring atom (an atom attached to exactly two other ring atoms) and "T" to indicate a fusion point or bridgehead atom with three or more bonds to other ring atoms. The "smallest set of smallest rings" definition is used to define a "ring" here, so that the "envelope" rings circumscribing smaller rings are not considered (Figure 4). There is also a TR screen that simply provides an indicator that an eight-membered or larger ring is present.

For the example structure, since the two six-membered rings are isolated (i.e., not part of larger ring systems), the screen used to describe them would be

TR 2 DDDDDD

Atom Count (AC). These screens are used to specify the minimum number of nonhydrogen atoms present in the graph (i.e., excluding any atoms described as single atom fragments).

The AC 17 screen would be used to describe the example structure, since the Cl atom of the hydrochloride is not included in the atom count.

Degree of Connectivity (DC). These screens specify the minimum number of atoms having at least a specified number of nonhydrogen atoms attached to them. DC screens are provided which specify nonhydrogen connectivities from three or more to six or more.

The example structure would be described by the screen

DC 6 3

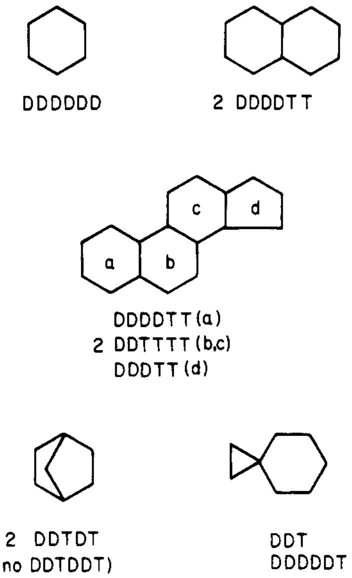


Figure 4. TR screen examples.

specifying the presence of six or more atoms having nonhydrogen connectivities of three or more.

Element Composition (EC). These screens specify the minimum number of atoms of each element (except H, D, and T) present in the connection graph, again excluding atoms described in single-atom fragments. The common elements have EC screens with specific counts, with the number of screens and counts depending upon the abundancy of the element in the CAS Registry file. Carbon, for example, has 14 EC screens covering a range of counts from 1 to 40, while oxygen has 12 screens, nitrogen has 8, and sulfur and chlorine each have 5. The less common elements such as gold or lead each have just one EC screen, specifying simply the presence of the element.

The screens used to describe the example structure, whose molecular formula is $C_{12}H_8ClNO_3{\cdot}HCl$, would be

EC	12 C	
EC	Cl	the Cl in the graph
EC	N	
EC	3 O	
GM	E Cl	the Cl in the single atom fragment

Graph Modifier (GM). Most GM screens are graph modifier element (GM E) screens that are used to describe the elements cited in the single-atom-fragment portion of the structure record.

For the example structure, the screen used to specify the presence of the hydrochloride salt would be

GM E Cl

Other screens are used to specify the following: (1) unusual structural features, e.g., unusual mass, valence, or charge attributes of atoms in the graph of the connection table or in the single-atom-fragment portion of the structure; (2) multicomponent substance data, e.g., the presence of two or more to four or more components and the presence of single atom fragments; (3) chemical substance class identifiers, e.g., classification of substances as alloys, incompletely described substances, minerals, mixtures, multicomponent substances, polymers, and radical ions, with further subclassification provided for incompletely described substances and polymers;[2] (4) text descriptor data, e.g., the type of stereochemical data provided by a substance's CAS Registry text descriptor.[18]

CAS ONLINE SCREEN DICTIONARY

The policy by which screens were selected for the CAS ONLINE screen dictionary has resulted in the presence of

Table III. Screen Distribution by Frequency

	screen no.		screens	
frequency range, %	count	% of total	count	% of total
0.00-1.00	701	33.1	3237	39.3
1.01-2.00	334	15.8	1330	16.1
2.01-3.00	215	10.1	1358	16.5
3.01-4.00	153	7.2	428	5.2
4.01-5.00	109	5.1	325	3.9
5.01-10.00	272	12.8	774	9.4
above 10.01	335	15.8	789	9.6

screens that have selectivity as well as chemical significance. Thus the searcher who is manually encoding a screen search can usually define the query with highly selective screens, i.e., screens that occur in a low percentage of the substances in the CAS Registry file. Almost 40% of the screen definitions in the dictionary are for structural fragments that occur in 1% or fewer of the structures on file, and more than 70% occur in 3% or fewer (Table III). (Screen number frequencies are somewhat different, due to screen number sharing, with 33% of the screen numbers referencing 1% or fewer of the structures on file and with 60% referencing 3% or fewer.)

The AA screen section of the dictionary, for example, contains almost 1000 screens with carbon as the central atom, with most including one or more heteroatoms among the attached atoms and many having bonds specified by type and value. Since carbon is the most common element in a file of organic substances, this depth of detail is necessary to get screens specific enough to be useful. Phosphorus, on the other hand, is relatively uncommon in the file substances, and only about 70 AA screens have it as their central atom, with only a few having bonds specified; most of these have frequencies under 1%, despite their lack of detail. The distribution of AA screens by central atom element and number of attached atoms is given in Table IV.

The HA and TW screens provide an ability to specify the presence of hydrogen atoms, needed because an AA screen that does not fully specify the attachments to the central atom implies that either hydrogen or nonhydrogen atoms could be present to complete valence requirements. Thus, the screen

HA C H -1 O -2 O

specifies a formyloxy group, while the screen

AA C -1 O -2 O

specifies simply any carboxylic ester group. The HA screens present in the CAS ONLINE dictionary are only those that occur relatively frequently, most having carbon, nitrogen, or oxygen central atoms. Only a few TW screens are present in the dictionary to describe common occurrences of CH_3, NH_2, OH, and SH groups.

The AS screens in the dictionary were selected[6] by generating all possible screens fitting chemically significant patterns, such as the patterns

R – C * C – R′
R – C * C * C – R′
R – C * C * C * C – R′

that describe 1,2-, 1,3-, and 1,4-disubstitution on a carbocycle. The frequency of appearance of each screen was then checked in a test file of structures. Screens that were overly specific were either deleted or grouped into related sets to share screen numbers, while screens that were too generic were made more specific by the addition of bond-type specifications. BS and CS screens were selected by similar procedures.[6]

The TR screens cover all possible combinations of fused and nonfused atoms forming three- to seven-membered rings. To increase the utility of these screens, additional entries were provided for frequently occurring rings with "or more" occurrence counts specified. The TR screen DDDDDD that describes an isolated six-membered ring, for example, has counts covering one to six occurrences, while the screen DDDDTT that describes an ortho-fused six-membered ring has counts covering one to four occurrences. In contrast, some infrequently occurring screens for three-, four-, and seven-membered rings were grouped into sets by screen number sharing.

Table IV. Augmented Atom (AA) Screen Distribution

	no. of AA screens				
		by no. of attachments			
central element	total	1	2	3	4
A	15	5		5	5
As	12	12			
B	16	16			
Br	1	1			
C	988	50	328	313	297
Cl	3	3			
F	3	3			
I	1	1			
M	33	33			
N	172	57	85	21	9
O	107	33	73	1	
P	72	39	13	5	15
S	78	33	27	9	9
Se	12	12			
Si	17	17			
Te	12	12			
X	15	15			
total	1557	342	526	354	335

The general structural feature screens (AC, EC, GM E, DC, and RC) were primarily chosen for comprehensive coverage. A basic "one or more" screen is provided for every general structural feature, with additional screens for higher counts provided where appropriate. Many of these screens are useful only for generic searches (e.g., all 14- to 16-carbon hydrocarbons) or for the NOT screen set used by automatic screen generation during a full-structure search. The screens most often used in an ordinary substructure search are the EC screens, used to specify the presence of an unusual element (e.g., Ge or Zn), and the GM E screens, used to specify a single-atom fragment.

Finally, the GM special purpose screens are provided to allow the searcher to select (or reject) substances in special classes (e.g., radical ions, coordination compounds, polymers, etc.) or to specify the presence of unusual structural features (e.g., charged atoms, isotopes, etc.).

SCREEN NUMBER SHARING

In the CAS ONLINE system, the number of unique screen numbers is determined by the size of the bit string used to implement the search procedure, since each screen number corresponds to a specific bit position. The size of the bit string is an important design factor: shorter bit strings mean faster searches, but longer bit strings mean more screens available for query encoding and thus a more versatile and more precise search system. The technique of assigning several screen definitions to a single screen number sidesteps a potential design dilemma by greatly increasing the number of screen definitions available to the user while only slightly increasing the size of the bit string. From another point of view, the technique reduces the number of unique screen numbers needed to handle a set of screen definitions, thus freeing some screen numbers for use elsewhere. The screen number sharing approach used for the EC and GM E screens for the less common elements, for example, handles 78 elements with only 24 screen numbers, saving 54 screen numbers for other uses.

Overall, screen number sharing allows the CAS ONLINE system to handle almost 6000 different screen definitions with only 2128 screen numbers.

Since the specification of a shared screen number in a search query would lead to the retrieval of structures containing any of the associated fragments, it might seem that this approach would degrade the selectivity of the search system. In practice, there is actually little or no loss of selectivity observed, since additional screens combined with AND logic can narrow down retrieval to the desired substructure. For example, the six possible AAs describing a carbon atom with two different halogen atom attachments are specified by screen 961, so the searcher would combine ("AND") 961 with additional AA screens to narrow down retrieval to a specific fragment; to select the bromo–chloro species, for example, the additional screens used would be

AA C Br

AA C Cl

Some irrelevant structures might still be retrieved that contain the desired screens but not the desired fragments; a structure containing the two structural fragments Br–C–I and C–Cl, for example, might be retrieved when the bromo–chloro species Br–C–Cl is sought. The number of such irrelevant retrievals will usually be quite low and is more than compensated for by the increased search power provided to the system and the searcher by the additional screens.

Several different schemes have been used for screen number sharing. First, there are many cases where several related screens will share a screen number. This is by far the most frequently used approach to screen number sharing and has been applied to many screens in the augmented atom and linear sequence classes, primarily to those of very high specificity. The typical screen set here contains screens with one or two "variable atoms", atoms that may have any element value in a set such as "halogen", "O or S", "N, O, S, or halogen", or "uncommon hetero (As, B, P, Se, Si, Te)". Other sets contain screens having different combinations of single, double, and normalized bonds.

Second, for most elements, no screen distinction is made between an atom of the element that appears in the connection table (described via an EC screen) and one that appears as a single-atom fragment (described via a GM E screen), since the corresponding EC and GM E screens have the same screen numbers. This was done because these elements occur so infrequently that the distinction between occurrences in the connection table and occurrences in single-atom fragments would be of little practical value. The distinction is made only for nine common elements (Br, Ca, Cl, H, I, K, N, Na, and O) that appear frequently enough as single-atom fragments that distinct GM E screens would be beneficial.

Third, the EC and GM E screens are assigned to screen numbers by using three different approaches. The most common elements, C and eight common heteroatoms (N, O, P, S, and halogen), along with As, B, Si, and 11 metals of interest have specific screen numbers assigned to them. Other elements are handled through a screen number sharing approach that assigns 27 screen numbers to groups and series of elements. A specific element can be selected by combining two screens with AND logic, one screen specifying a vertical group from the periodic table and the other screen a horizontal series. Silver, for example, would be pinpointed by screens 1920 AND 1921 (specifying group 1B and transition series 2, respectively), and 45 other elements are similarly accessed. The 33 least common elements, such as Ac, cannot be accessed specifically, but only via one of the generic screens specifying a group or series.

Fourth, a number of BS screens are organized into several groups of sets of related screens, with each specific BS fragment appearing in two sets of screens. One group of screen sets, for example, describes two rings linked by a short chain; some sets consist of fragments with the same ring bonds but different combinations of chain bonds, and the other sets consist of fragments with the same chain bonds and differing ring bonds. Screen 834, for example, specifies two rings linked by three −1 bonds, while 845 specifies two rings with *4 bonds linked by different combinations of three chain bonds; the fragment

BS A *4 A −1 A −1 A −1 A *4 A

(describing two alternating bond rings connected by a chain of three single bonds) appears in both screen sets and would be pinpointed by combining them with Boolean logic as 834 AND 845.

Finally, there is an implicit screen number sharing accomplished through the use of generic element symbols in AA screen definitions, with "A" representing any atom, "M" any metal, and "X" any halogen.

FRAGMENT GENERATION

In the CAS ONLINE file-building procedure, the creation of substructure search screens for a substance begins with a fragment generation step, in which structure fragments and other search data are generated from the Registry connection table of the substance. The generation procedure is exhaustive, creating all possible structure fragments. A subsequent screen generation step checks the generated fragments against a dictionary. The presence of structure fragments that are cited in the dictionary is recorded in a bit string that is the screen search file record. Fragments not cited in the dictionary are ignored; such fragments are, for the most part, very frequently occurring and thus of little selectivity or else are so rare that they would not be used enough in searching to justify their inclusion in a dictionary of limited size.

Most of the structure fragments are generated from what is termed the graph of the connection table, i.e., the atoms and bonds that comprise the basic structure of the substance. Simple derivatives of this basic structure (i.e., hydrates and simple metal or acid salts) that are described via the single-atom-fragment portion of the Registry record lead to the generation of GM E fragments.

As stated above, the fragment generation procedure generates all possible fragments. While this might seem inefficient, it is both easier and faster than would be the generation of only those fragments that appeared in the screen dictionary. The augmented atom generation procedure, for example, generates AA fragments for each nonhydrogen atom in the graph by using a procedure that considers all possible combinations of attached atoms and three levels of bond specification (none, bond type only, and bond type and value). For the carboxy group carbon (atom d) of the example structure shown earlier, the procedure would begin with the most specific fragment

AA C −1 C −4 O −4 O

and generate 14 additional less-specific fragments (Figure 5). (The procedure maintains a list of fragments, checking each generated fragment against the list; if the fragment is new, it is added to the list with a count of one, while if it is already present, the appropriate counter is incremented.) Subsequently, the screen generation step would check the screen dictionary for each fragment and would find only six of them, those shown underlined, to be present; the bits for these fragments would be set in the screen search record bit string.

A search screen includes a count or number that gives the number of times that the fragment appears in the structure. The count is a minimum value, not an exact value. The "or

Starting fragment:
AA C −1 C −4 O −4 O

Additional generated fragments:

AA C C	AA C − C − O − O
AA C − C	AA C O
AA C −1 C	AA C − O
AA C C O	AA C −4 O
AA C − C − O	AA C O O
AA C −1 C −4 O	AA C − O − O
AA C C O O	AA C −4 O −4 O

Figure 5. Example of fragment generation.

more" ability is implemented during the screen generation step that creates the bit string used in searching. If the structure contained five oxygen atoms, for example, the fragment generation step would create the fragment

EC 5 O

The screen generation step would set the bits for fragments with a count of five or less

EC O through EC 5 O

so that the structure would be retrieved, for example, by a search query specifying that three or more oxygen atoms must be present. This approach, like the "inclusive" approach used in AA and other screen definitions, facilitates substructure searching by automatically providing for the retrieval of larger structures that contain the desired substructure.

MANUAL SCREEN SEARCH QUERY ENCODING

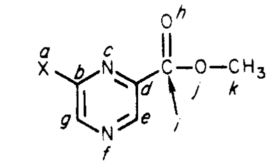
When a screen search query profile is being encoded, most key structural fragments defining the substructure sought can be identified by concentrating on two aspects of the query: those portions of the substructure that are fully defined and those that are unusual (e.g., containing several heteroatoms, uncommon elements, several multiple bonds, or uncommon ring sizes).

In general, AA and AS screens are the most precise screens, and effective screen profiles can be developed by using only these two types of screens. The searcher should concentrate on AA and AS fragments that contain several heteroatoms or unusual bonding patterns, initially looking at larger fragments and longer sequences. The fragments most likely to give useful screens include the following: (1) AA fragments with carbon as a central atom and several attached heteroatoms, with a hetero central atom, or with carbon as a central atom and three or four attached atoms (describing a ring-fusion point or bridgehead atom, a ring atom with one or two acyclic substituents, or a chain branch point); (2) AS sequences containing several heteroatoms or containing both ring and chain bonds (describing a chain–ring–chain or ring–chain–ring path).

The other types of screens are less generally useful but can still provide good selectivity when they are applicable. For example, HA and TW screens are useful when the presence of hydrogen on an atom must be specified, although the selection of screens in the screen dictionary is a bit limited. TR screens are primarily useful when describing isolated rings or systems that are large polycyclic ring systems, rings with three, four, seven, or eight or more nodes, or spiro systems. BS screens are useful when the query structure contains chain–ring–chain or ring–chain–ring paths with specified bond values or paths containing multiple bonds. CS screens, because of their exact specification of connectivity values, are primarily useful when describing a fully defined portion of a substructure (i.e., where additional substituents cannot occur or are not desired) or when performing full-structure searches.

The special GM screens are useful when the searcher wishes to select (or reject) particular classes of substances as answers. These screens allow access to the various substance class identifiers (e.g., polymer, radical ion) used in the CAS Registry system. The type of stereochemical data provided by substances' text descriptors may also be specified, although a familiarity with CAS stereochemistry practices is required to use these screens effectively.[18]

Query:

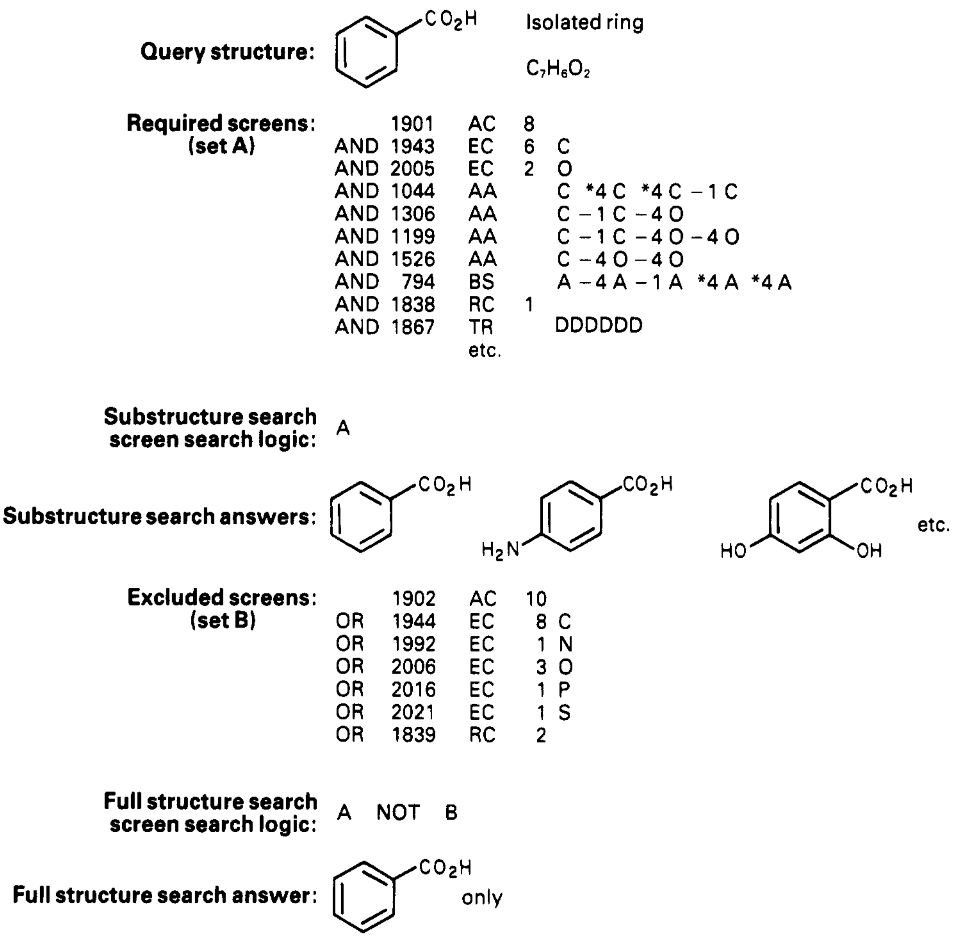


*X=F,Cl,Br or I
*isolated ring

	Screen Number			Screen Definition	Freq.	Atoms Described	Notes
	1197	AA		C* C* N − X	0.65%	b	(1)
AND	1589	AA	2	N*4C*4C	4.88%	c,f	
AND	1122	AA		C*4C −1C*4N	2.48%	d	
AND	1224	AA	3	C*4C*4N	2.48%	b,d,e,g	(2)
AND	1523	AA		C −1O −2O	16.38%	i	
AND	1506	HA		CH3 −1O	16.30%	k	
AND	212	AS		hal C C N C	2.92%	a-b-g-f-e	(3)
AND	314	AS		hal − C*C*N	0.97%	a-b-g-f	(1,3)
AND	401	AS		O − C − C*N	3.50%	j-i-d-c	
AND	371	AS		O − C − C*C*N	2.66%	j-i-d-e-f	
AND	382	AS		N*C*C*N	8.68%	c-d-e-f	
AND	268	AS		C − O − C − C*N	1.15%	k-j-i-d-c	
AND	881	BS		A −1A*4A*4A −1A −1A	10.29%	a-b-c-d-i-j	
AND	882	BS		A −1A*4A*4A −1A −2A	5.66%	a-b-c-d-i-h	
AND	1867	TR		DDDDDD	53.29%	b-c-d-e-f-g-b	

Notes: (1) These two screens are the key screens.
(2) A screen with a count of 4 is not available in the dictionary.
(3) These screens are actually sets of four screens with the same screen number, with hal = F, C1, Br and I.

Figure 6. Typical manually encoded screen search.



Required screens: (set A)

	1901	AC	8	
AND	1943	EC	6	C
AND	2005	EC	2	O
AND	1044	AA		C *4 C *4 C -1 C
AND	1306	AA		C -1 C -4 O
AND	1199	AA		C -1 C -4 O -4 O
AND	1526	AA		C -4 O -4 O
AND	794	BS		A -4 A -1 A *4 A *4 A
AND	1838	RC	1	
AND	1867	TR		DDDDDD
		etc.		

Substructure search screen search logic: A

Excluded screens: (set B)

	1902	AC	10	
OR	1944	EC	8	C
OR	1992	EC	1	N
OR	2006	EC	3	O
OR	2016	EC	1	P
OR	2021	EC	1	S
OR	1839	RC	2	

Full structure search screen search logic: A NOT B

Figure 7. Example of query encoding.

A typical manually encoded screen search is shown in Figure 6. The screen frequencies are for their appearance in a uniform 1% sample of the full CAS Registry file and are shown to indicate the relative utilities of the screens. The screens selected are not all of the possible screens but only those that would provide useful selectivity. Many redundant screens, such as

EC 2 N
EC 2 O
AA C – O – O

have been omitted; they would add nothing to the selectivity of the search, since their presence is implied by other screens that were used. Note that the most of the selectivity of the search is provided by the two key screens 1197 and 314, describing the halo-substituted ring; the remaining screens further describe the desired substructure but provide only a little additional specificity.

In the initial version of CAS ONLINE, the search capability was limited to screen searching. While screen searches could usually provide satisfactory precision for typical substructure search queries, they did on occasion suffer from the problem of "false drops", answer structures containing the specified structural fragments in other than the desired relationship. A further drawback was that query encoding, though not difficult, was at times tedious. These faults were eliminated with the introduction of graphic structure input and atom-by-atom searching in late 1981.

AUTOMATIC SCREEN SEARCH QUERY ENCODING

In the present CAS ONLINE system, screen searches are still used, but the searcher is usually not directly involved (unless special GM screens have been specified). Instead, after a search query has been input graphically as a structure diagram, the system automatically generates screens defining the query by using a procedure analogous to fragment generation. The generated screens that appear in the system dictionary are then used to construct a screen search query. For a full-structure query, the system also generates a set of screens that must *not* be present, so as to take advantage of the restrictive nature of this type of query (see Figure 7). Finally, answers from the screen search are routed to the atom-by-atom search procedure, and only those matching the search query definition are routed to the final answer file and the searcher.

Although the emphasis is now on query definition via structure diagrams, CAS ONLINE will continue to provide for manually encoded screen searches. Some types of generic searches such as "14- to 16-carbon hydrocarbons" or "pentacyclic substances containing a three-membered ring" cannot be done efficiently through query structure input but are easily handled by manually encoded screen searches.

ACKNOWLEDGMENT

The development of the CAS Chemical Registry system was substantially supported by the National Science Foundation. Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support. CAS also acknowledges the substantial technical contribution made by the Basel Information Center for Chemistry in the development of the screen set used in CAS ONLINE. Significant contributions were made to the design of the initial version of the automatic screen generation procedure by D. Lefkovitz and H. Hill of the University of Pennsylvania.

REFERENCES AND NOTES

(1) Farmer, N. A.; O'Hara, M. P. "CAS ONLINE—A New Source of Substance Information from Chemical Abstracts Service". *Database* **1980**, *3*, 10–25.

(2) Dittmar, P. G.; Stobaugh, R. E.; Wilson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.

(3) Wigington, R. L. "Machine Methods for Accessing Chemical Abstracts Service Information". In "Proceedings of IBM Symposium on Computers and Chemistry"; IBM Data Processing Division: White Plains, NY, 1969.

(4) Richman, S.; Hazard, G. F., Jr.; Kalikow, A. K. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program". In "Retrieval of Medicinal Chemical Information". *ACS Symp. Seri.* **1978**, *No. 84.*

(5) Schenk, H. R.; Wegmuller, F. "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 153–161.

(6) Graf, W.; Kaindl, H. K.; Kniess, H.; Schmidt, B.; Warszawski, R. "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 51–55.

(7) Graf, W.; Kaindl, H. K.; Kniess, H.; Warszawski, R. "The Third BASIC Fragment Dictionary". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177–181.

(8) Farmer, N. A. "The Proposed Chemical Abstracts Service's Substructure Search System". In "Proceedings of the Technical Information Retrieval Committee of the Manufacturing Chemists Association"; Arlington, VA, Aug 1977; McNulty, P. J., Smith, R. B., Eds.; Manufacturing Chemists Association: Washington, DC, 1977.

(9) Zeidner, C. R.; Amoss, J. O.; Haines, R. C. "The CAS ONLINE Architecture for Substructure Searching". In "Proceedings of the 3rd National Online Meeting"; Learned Information, Inc.: Medford, NJ, 1982; pp 575–586.

(10) Hagadone, T. R.; Howe, W. J. "Molecular Substructure Searching: Minicomputer-Based Query Execution". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 182–186.

(11) Blake, J. E.; Farmer, N. A.; Haines, R. C. "An Interactive Computer Graphics System for Processing Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 223–228.

(12) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186–192.

(13) Feldman, A.; Hodes, L. "An Efficient Design for Chemical Structure Searching. I. The Screens". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147–152.

(14) Lynch, M. F. "Screening Large Chemical Files". In "Chemical Information Systems"; Ellis Horwood: Chichester, 1975.

(15) Dunn, R. G.; Fisanick, W.; Zamora, A. "A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212–218.

(16) Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VI. Substance-Related Statistics". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76–82.

(17) Mockus, J.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18–22.

(18) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3–8.

DARC Substructure Search System: A New Approach to Chemical Information[†]

ROGER ATTIAS

Association pour la Recherche et le Dêveloppement en Informatique Chimique (ARDIC), 25 Rue Jussieu,
75005 Paris, France

Received November 29, 1982

The efficiency of a chemical information system depends upon information parameters retained by the language used to describe compounds. Structural-based languages provide a specific approach to chemical problems. The DARC system allows a coherent approach to substructure search, structure–activity correlation, and computer-aided design by defining relationships between the notions of substructure, structure, and family of structures. The substructure search is based on the concept of fuzziness: it is expressed in terms of subgraph isomorphism between a set of fuzzy graphs and a set of graphs. The file to be searched is processed by an automat which generates multilevel fuzzy graphs corresponding to local descriptions of the defined structures. The DARC descriptions of these graphs are stored in a tree structure. The same process is applied to the fuzzy graph of a query. As a result of this approach, the user language is the natural language of the chemist: free drawing of the substructural diagram with no use of a dictionary for the search. The retrieved structures can be displayed on a graphic terminal. These principles have been applied to the full CAS Registry Structure File and have made possible, for the first time, on-line substructure searches on 5 million compounds (EURECAS). An automatic link to the textual data base (CA SEARCH) makes it possible to deal with both structural and textual aspects of the query.

INTRODUCTION

The DARC system and its role in French national computer science policy were presented in 1972 by Professor Jacques-Emile Dubois.[1] This system, developed since 1963,[2–6] places chemical compounds in their structural context and accounts for their local and global properties by using their topology as a starting point.[7,8]

Structural information is handled so as to achieve a coherent approach to the notion of substructure, structure, and family of structures (hyperstructures).

The substructure is perceived as a generalization of the notion of structure. The substructure search system, which is an application of the basic principles, constitutes the first and necessary step forward in handling chemical problems. One aim of this system was to make possible access to Chemical Abstracts Service (CAS) products, not only by texts but also by structures through a structural user language reflecting the thought of the chemist and enlarging his field of investigation. To achieve these goals, our constraint was to perform a purely topological approach, avoiding, whenever possible, any type of global fragmentation. The first tests were on samples;[8,9] the difficulties then perceived as to volume and response time were gradually solved and led to on-line search in 1978 of CBAC and in 1980 of EURECAS (commercialized in Feb 1981).

In this paper we introduce the different steps of our research, its context, its results, the general methodology which has guided our approach, and its place in the evolvement of structural languages: the concepts and the technical aspects will be more fully developed in a series of papers in which, in particular, will be discussed the important contributions of

[†] Presented at the 182nd National Meeting of the American Chemical Society, New York, August 1981.