A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature†

RONALD G. DUNN, WILLIAM FISANICK,* and ANTONIO ZAMORA
Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received June 20, 1977

In January 1976 Chemical Abstracts Service (CAS) began operation of an experimental chemical substance search service which offers both retrospective and current awareness searches for specific substances and for substances containing specified substructures. To support this service, CAS has developed an experimental computer-based substructure search system which provides for batch mode serial searches on files containing *Chemical Abstracts* Index Nomenclature and structurally related data. The search system supplements basic text search methods with screening techniques to improve search efficiency and with extended logic capabilities to improve the precision of search results.

INTRODUCTION

Chemical Abstracts Service (CAS) builds and maintains a large computer-readable information base[1] to facilitate the production of *Chemical Abstracts* (CA) and other CAS publications and services. The CAS information base is comprised of several files which contain bibliographic information, abstract text, and data for the CA indexes. In addition to this document-oriented information, CAS maintains chemical substance authority files to support production operations; among these are the Chemical Structure and Nomenclature Files of the CAS Chemical Registry System.[2]

For several years CAS has been exploring various techniques for performing substructure or compound class searches on the information in the substance authority files. Experiments with topologically based substructure searching began in 1963.[3,4] Topological substructure searching uses iterative procedures for matching atom-by-atom the computer representations, such as connection tables, of substructure queries against equivalent representations of specific substances on a file. By 1968 CAS had a comprehensive set of substructure search programs[16] which utilized the CAS Chemical Registry System files.[6] However, the generation of topological substructure screens required a considerable amount of computer resources and the searching process required highly specialized computer programs and search files. Furthermore, the search technique and files did not provide for the direct correlation of the retrieved substances with the corresponding textual information on their uses, activities, and other properties in the CAS data base. Consequently, CAS turned its efforts toward developing techniques for nomenclature-based searching as a possible alternative to topological substructure searching. Research in this area has led to the use of standard text search techniques to perform substructure searches based on the structural information contained in CA Index Names. Since files containing CA Index Names, such as "CA Subject Index Alert" and "Chemical-Biological Activities", also contain corresponding descriptive information, direct correlation of substance information with related nonsubstance information (i.e., substructure-text searching) is possible using text search techniques. To support the encoding of effective nomenclature-based substructure search profiles, a series of search manuals and other support material has been developed.[7]

Substructure and substructure-text searching based on CA Index Nomenclature has been demonstrated successfully using small files[7,8] and using on-line retrieval files such as CHEMLINE, which is available in the toxicology information service of the National Library of Medicine[9,10] and, more recently, CHEMNAME, which is available in Lockheed's DIALOG information service.[11] Also, the Systems Development Corporation has recently announced plans to add "CA Subject Index Alert" data, including the CA Index Names, to their ORBIT system.[12] The substructure search system described in this paper is designed to extend the basic nomenclature search capability by permitting effective searching of very large files. This system is used by CAS to support an experimental chemical substance retrieval service.

SEARCH SERVICE

The experimental substance retrieval service allows the CAS database to be searched for CA references to chemical compounds. The service, which has been offered since January 1976, is properly called "Compound Searches, an Experimental Service". Through this service, searches are performed at CAS for specific chemical compounds reported in the period 1907 to date. Searches for classes of chemical compounds which possess certain defined structural features (i.e., substructure searches) are performed in the period 1967 to date. An alerting service is also available to retrieve citations either for specific chemical compounds or for classes of structurally related compounds as soon as those citations are entered into the database. Manual search techniques are used for specific substance searches in the period 1907 to 1967. Substructure searches from 1967 to date are conducted by computer using the substructure search system.

For substructure search requests processed through the experimental service, query profiles are developed by CAS staff according to the specifications submitted by the customer. The profiles are used to conduct test searches on a sample file. Normally, the profiles used for the test searches are designed to permit the retrieval of some substances that have structural features which vary slightly from the initial query specifications. The use of such generic profiles gives customers an opportunity to modify their original query specifications to ensure retrieval of all the potentially relevant substance information from the database. The profiles are then refined, if necessary, and all or part of the retrospective search database is searched, depending on customer requirements.

The retrieval supplied to the customer for a substructure search query consists of the Chemical Substance Index entries and related data for the substances in the specified structural class. Two reports are generated by the system for each search query. One report (see Figure 1) is ordered by CAS Registry Number and brings together all of the retrieved information for each chemical substance which satisfies the profile requirements. This report includes CAS Registry Numbers, CA Index Names, ring descriptions (when present), and CA reference numbers. Many CA reference numbers have associated textual descriptions which indicate the types of in-

† Presented, in part, to the Division of Chemical Information, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 31, 1976.
* To whom correspondence should be addressed.

SEARCH OUTPUT - I

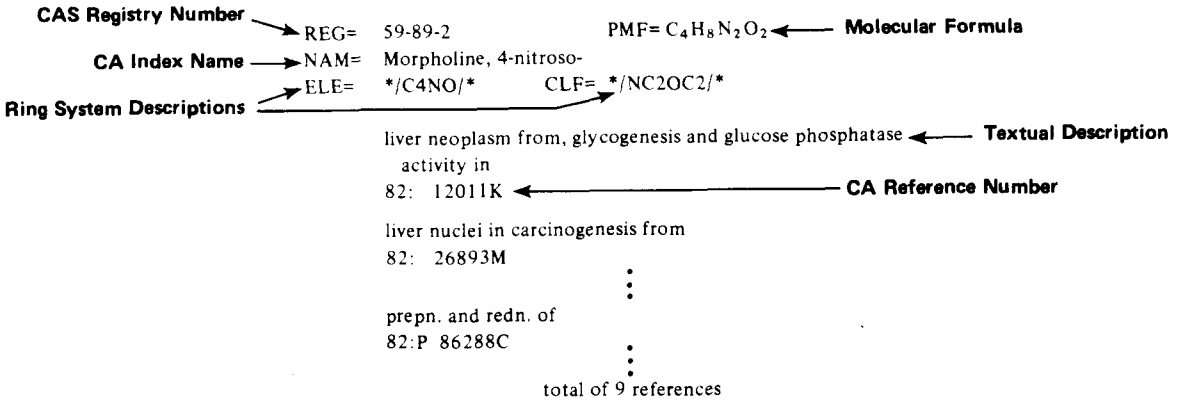


Figure 1. Printout in CA Registry Number order.

SEARCH OUTPUT - II

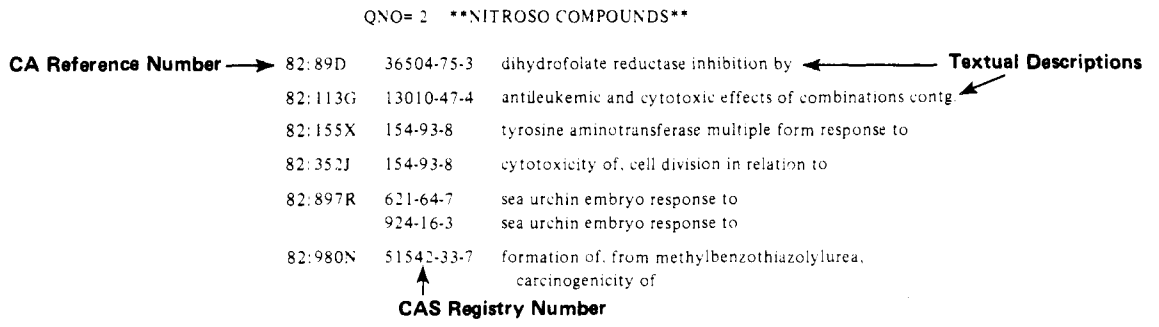


Figure 2. Printout in CA reference number order.

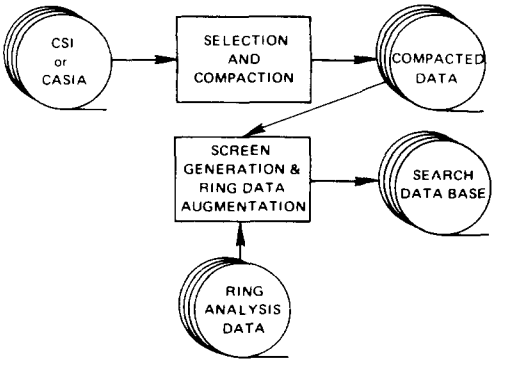
formation (e.g., properties or uses) which can be found in the corresponding original articles. These textual descriptions permit the user of the service to do some further screening (manually or automatically) if the user is interested in a particular concept associated with a class of substances. For example, the retrieval shown in Figure 1 is from a search for nitroso compounds. If the interest is in the relationship of nitroso substances with liver cancer, then the first two references would probably be pertinent, but, not the third one, in which the main emphasis is the preparation and reduction of 4-nitrosomorpholine.

The retrieval shown in Figure 1 also illustrates how difficult it would be to locate a class of substances such as nitroso compounds in the printed, alphabetically ordered CA Chemical Substance Index. Essentially, it would be necessary to scan all of the entries of the index since the "nitroso" term is not used as an index entry point according to CA Index Nomenclature rules. However, in a text or character-matching search of computer-readable nomenclature files, the "nitroso" string can be easily located and the corresponding CA Index Names retrieved.

The second report (see Figure 2) is arranged by CA reference number order and provides a convenient means for retrieving further information from the printed or microform issues of CA. For each CA reference cited in this report, the retrieved CAS Registry Numbers and their associated textual descriptors are included, when such descriptors are present.

SEARCH DATABASE

A general flowchart of the file-building operation used to create the search file is shown in Figure 3. Two versions of the search file are built, one for retrospective searching and one for alerting or current awareness searching. The retrospective file is derived from files used to produce the printed CA Volume Chemical Substance Indexes (CSI) and the



CSI = Chemical Substance Index (retrospective)
CASIA = Chemical Abstracts Subject Index Alert (current awareness)

Figure 3. Generation of the search database.

Parent Compound Handbook.[13] The Alerting Service search file is produced by extracting the chemical substance index entries from "CA Subject Index Alert" (CASIA), a biweekly computer readable service offered by CAS.

In the file-building operation, the substance data are selected, compacted, organized into search file format, and augmented with substructure bit screens and ring descriptors (retrospective file only). The content of the substructure bit screen is described in the following section.

The format of the search file is such that only a single search is required to retrieve both the substance and reference information. This format is analogous to the retrieval format shown in Figure 1. For each substance, there is an initial substance-record followed by one or more reference-records. The link among the records is the CAS Registry Number. Each substance-record contains a Registry Number, CA Index Name, molecular formula, substructure bit screen, and ring descriptor, when appropriate. Each reference-record contains

Molecular Formula Screen	Ring Screen	Nomenclature Screen
30 bytes	20 bytes	9 bytes

Figure 4. Substructure bit screen.

the CA reference number, Registry Number, and textual descriptors for the CA reference, if such descriptors are present. In searching the database, each substance-record is examined, and if the substance satisfies the query specifications, the substance-record and its corresponding reference-records are output to the retrieval file.

The retrospective file is segmented on a volume basis, with each volume of the search file corresponding to a volume of CA. As of November 1976, CAS has created 19 subfiles containing the chemical substance index entries for Volumes 66–84 of *Chemical Abstracts* covering the years from 1967 to the first half of 1976. Any specific volume or set of volumes can be searched individually. When more than one volume is searched, the retrievals for individual substances from different volumes are merged in the output. The total retrospective file is very large. There are approximately 4.2 million substance-records and 9 million reference-records for a total of over 13 million records on the file. The number of unique substances on the file, approximately 2.8 million,[17] is considerably less than the number of substance-records because the same substance can be cited in more than one volume. The search file currently resides on 72 reels to magnetic tape (1600 bpi), although because of volume segmentation not all of the reels are full reels.

Issues of the current awareness database are generated every two weeks corresponding to issues of CASIA. The current awareness database differs from the retrospective database in that the ring descriptions are not included and the file is not sorted into Registry Number sequence.

SUBSTRUCTURE BIT SCREEN

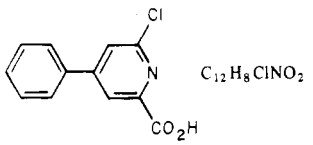
The purpose of the substructure bit screen is to determine in an efficient manner if a substance contains certain generic structural features. It is, in essence, a series of "YES/NO" flags which can be rapidly scanned by computer to determine whether specified structural features are present or absent in the substance-record. Only if a substance satisfies this screen is the more precise but more time-consuming text searching performed on that record. The substructure bit screens are derived from the molecular formulas, from ring features, and from the CA Index Names of chemical substances. These partial bit screens are then juxtaposed to form for each substance a single bit screen containing 472 bit positions (see Figure 4). The composite bit screen makes it possible to combine nomenclature and structural features for more efficient searching. A more detailed description of the bit screens is given below.

Molecular Formula Screen. Bit screens derived from molecular formulas summarize the elemental composition of a substance by indicating which chemical elements are present. For certain commonly occurring elements, such as chlorine, the number of atoms present in the substance is also recorded. Thus, if a substance contains four chlorine atoms, the bit screen indicates that one, two, three, or four chlorine atoms may be found in the substance. This makes it possible to retrieve such a substance as a potential answer to a query requiring a minimum of three chlorine atoms using a single logical operation.

The presence and number of some important groups of elements are also recorded in the molecular formula screen. Some of the groups include halogens, chalcogens, metals, transition series, and other periodic table groups.

Letter	Letter Value	Digram	Bit
F	22	FO	36 = (64 x 22 + 31) Modulo 61
O	31	OR	5 = (64 x 31 + 34) Modulo 61
R	34	RM	9 etc.
M	29	MA	43
A	17	AL	18
L	28	LD	43
D	20	DE	20
E	21	EH	26
H	24	HY	52
Y	41	YD	21
D	20	DE	20
E	21	–	–

Figure 5. Derivation of a nomenclature screen.



2-Pyridinecarboxylic acid, 6-chloro-4-phenyl-

Figure 6. Example of bit screen generation.

Ring Screen. The portion of the substructure bit screen which indicates the presence of ring features in a substance is derived from the elemental compositions of the ring systems. Seventy-five bits correspond to the specific ring systems which account for approximately 95% of all ring system occurrences in the substances recorded in the CAS Chemical Registry System. Additional bits are used to denote the individual component rings of larger ring systems, the number of rings present, the ring sizes, and generic ring information (e.g., presence of a carbocyclic ring system).

Nomenclature Screen. The nomenclature bit screen for a given substance is derived from the alphabetic digrams which are present in the CA Index Name for that substance. (For example, the term "ethyl" contains the digrams ET, TH, HY, and YL.) The CA Index Name is analyzed by a computer algorithm which identifies the digrams that are present and converts them to a short, coded representation which is stored in the nomenclature bit screen. Because of this encoding process, a single bit in the nomenclature screen may correspond to more than one alphabetic digram. However, rapid computer scanning of the nomenclature bit screen makes it possible to determine in an efficient manner whether the corresponding substance name is likely to satisfy the nomenclature requirements of the search profile.

The nomenclature bit screen is derived by using a hashing technique similar to that suggested by Harrison.[14] Each nomenclature screen consists of 61 bits. The bit corresponding to each alphabetic digram is given by $(64\alpha + \beta)$ [modulo 61] wherein the alphabetic characters A through Z take the consecutive values from 17 to 42, respectively. Figure 5 illustrates the derivation of the nomenclature screen for the term "formaldehyde".

Screen Generation. The types of screens that are generated for the retrospective database will be illustrated for the substance shown in Figure 6. The CA Index Name for the substance is "2-Pyridinecarboxylic acid, 6-chloro-4-phenyl-".

The molecular formula subscreen for this substance includes bits corresponding to the presence of at least eleven carbon atoms, five hydrogens, one chlorine, one nitrogen, and two oxygens, and also of one halogen atom, two chalcogens, and three NOPS, i.e., three atoms from the nitrogen, oxygen, phosphorus, and sulfur group.

The ring subscreen records the presence of a C_6 system, a C_5N system, two unique ring systems, a 1-ring system, a

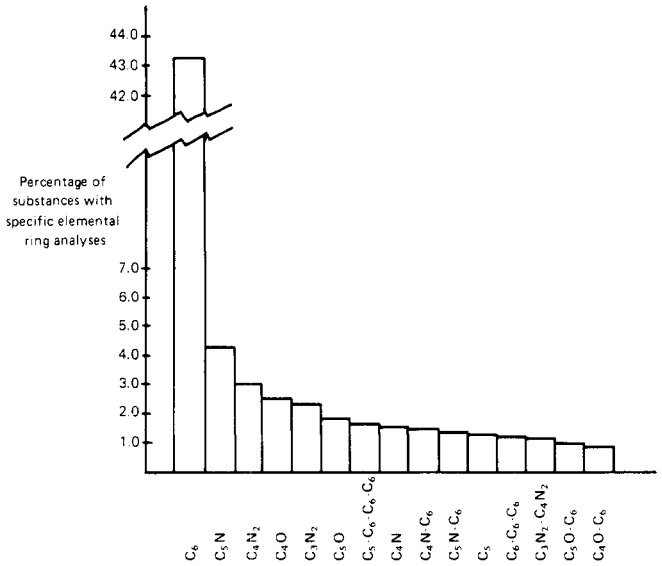


Figure 7. Fifteen most common elemental ring analyses in Volume 82 substances.

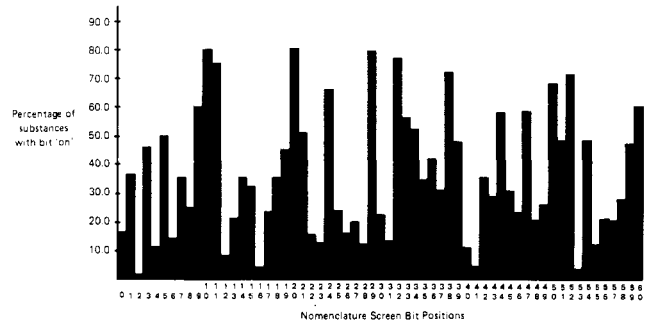


Figure 8. Nomenclature screen characteristics.

6-atom monocycle, a nitrogen heterocycle, and a carbocycle.

The nomenclature subscreen records the probable presence of the digrams PY, YR, RI, ID, DI, etc., in the Index Name.

To determine some of the characteristics of the screens generated from the retrospective search file, we obtained statistics on the Volume 82 portion of this file. This volume contains 265 508 unique substance-records. Of 236 possible molecular formula characteristics, only 38 occurred in more than 20% of the substances. Most of these 38 characteristics were counts of commonly occurring elements. Only carbon, hydrogen, nitrogen, and oxygen occurred in more than 20% of the substances. Consequently, in substructure queries that contain elements other than carbon, hydrogen, nitrogen, and oxygen, a file screenout greater than 80% should be obtained by using only molecular formula screens.

Table I. Ring Characteristics Occurring in More Than 10% of Volume 82 Substances

Description	% Substances
1. One or more unique ring systems	72.5
2. Six-atom component ring	65.3
3. C_6 component ring	58.1
4. One-ring isolated system	56.4
5. Carbocycle	51.7
6. Six-atom monocycle	49.4
7. C_6 isolated ring system	43.2
8. Heterocycle	39.6
9. N-heterocycle	29.0
10. Five-atom component ring	27.2
11. Two or more unique ring systems	24.8
12. Two-ring isolated system	16.9
13. O-heterocycle	14.4
14. "Other" isolated ring system	12.5
15. Five-atom monocycle	12.0

The ring characteristics present in more than 10% of the Volume 82 substances are listed in Table I. These screens represent about 10% of the total number of ring screens, i.e., 15 out of 151, and are primarily generic characteristics. The C_6 ring (numbers 3 and 7) is the only specific ring that occurs in more than 10% of the substances. A file screenout greater than 90% should be obtained for substructure queries that contain non-C_6 rings, either isolated or fused, i.e., embedded in a larger system. The 15 most common elemental ring analyses of ring systems in Volume 82 substances are shown in Figure 7.

The characteristics of the nomenclature screen for Volume 82 substances are summarized in Figure 8. The two bit positions with the highest number of postings are 20 and 10. Bit 20, which is posted for 79.4% of the substances, represents the digrams, AN, BK, CH, DE, EB, RX, SU, TR, UO, VL, WI, XF, and YC. Bit 10 (79.2%) represents DA, BA, NZ, OW, PT, QQ, RN, TH, and UE. The posting frequency of bit positions and, consequently, their effectiveness as screens depends on the frequency of occurrence of the corresponding digrams in CA Index Names. We obtained digram statistics on the Index Names in Volume 77 of the CA Chemical Substance Index and found that 25 digrams accounted for more than 50% of all digrams in the names (see Figure 9).

Although any single digram may not screen out a substantial portion of the search files, the use of multiple digrams is highly effective in selecting the substances for which further nomenclature searching is required. For example, in one ex-

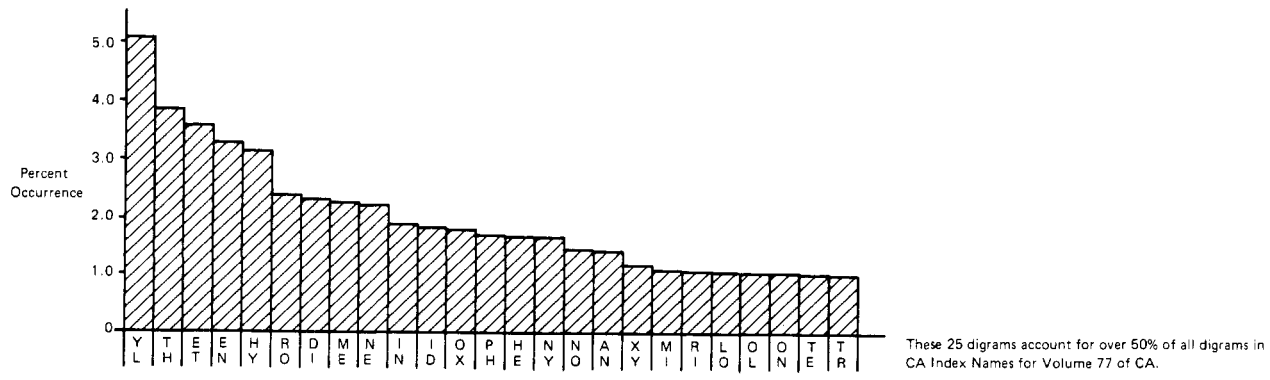


Figure 9. Chemical nomenclature characteristics.

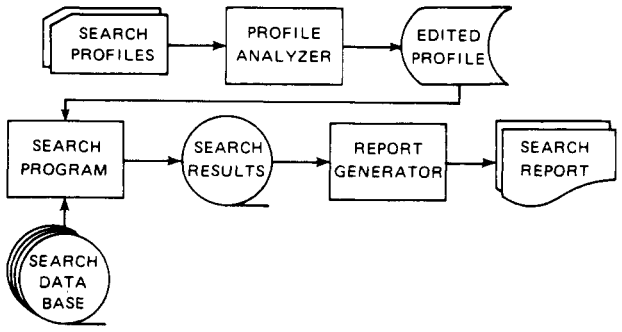


Figure 10. Search operation.

periment, using only the nomenclature screens, a search conducted of a batch of 11 substructure search queries containing a variety of query types yielded an average query screenout of 91.9%.

Generation of the substructure bit screen from the nomenclature and supporting data is a very rapid process relative to some other substructure screen generation processes that have been used, mainly because the generation process is primarily a tabulation of existing data.

SEARCH SYSTEM DESIGN

The search system consists of three major programs: a profile analyzer, a search program, and a report generator. The programs are written in PL/1 (optimizing complier) and use several assembler language subroutines. Figure 10 illustrates the dataflow through these programs.

The basic function of the profile analyzer is to organize the profile so that it can be searched efficiently. The input consists of a set of profiles which are derived by a chemical information specialist from the substructure queries. Profiles are developed using search aids which relate structural features to appropriate nomenclature search terms. One of the specific functions of the profile analyzer is to detect encoding errors. Other important functions include converting the search terms to the proper character storage mode, constructing substructure bit screens for the queries, printing the profiles for review by the encoder, and organizing the profiles into search tables for processing by the search program.

The system's search program is basically a text search program, but additional features have been added to improve the search efficiency and the precision of the retrieval. To improve search efficiency, a substructure screen search capability has been added. This screen search reduces the number of substance-records which require the more time-consuming text search. To improve the precision of the retrieval, some special logic capabilities have been added which take advantage of the rigid format of CA Index Nomenclature. These additional capabilities are discussed in more detail below.

The search program allows searching a single query or a set of queries, with the query profiles residing in computer memory at search time. The storage for the profile is allocated dynamically so there are no restrictions on the number of search terms or the number of queries. The largest search run at CAS thus far was a batch of 17 queries that had over 1000 search terms. This search required 436K of computer memory while typical searches require only 200–300K of computer memory.

The results of the search program are processed by the report generator to produce the final output of the system. The report generator separates the results of each query and prints the two reports discussed above (see Figures 1 and 2).

CAPABILITIES OF THE SEARCH PROGRAM

The system's search program has three distinct levels of search with a separate profile being used for each level. Some

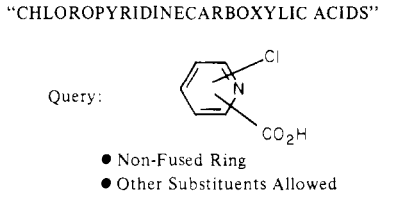


Figure 11. Sample substructure search query.

"CHLOROPYRIDINECARBOXYLIC ACIDS"

Molecular Formula=C6H1CL1N1O2
Number of Rings=1
Ring System Size=6
Ring System Analysis=C5N
Name Digrams=CHLORO CARBOXY PYRIDIN

Figure 12. Screen search profile.

specific characteristics of these search levels will be illustrated, using the same query shown in Figure 11.

Screen Search. The first level of the program, screen search, performs a search on the substructure bit screens to reduce quickly the number of substance-records that require further, more time-consuming searching.

The profile analyzer automatically converts the information from the screen search profile into a substructure bit screen using the same routines that generate the screens for the search file. At search time, the search program compares the profile bit screens with all the bit screens on the search file to identify the substance-records that contain the required attributes. All screen search profile requirements are expressed as minimum requirements to ensure retrieval of all relevant substance-records. For example, the screen search profile for the sample query (see Figure 12) will retrieve substances that have a molecular formula with *at least* six carbons, one hydrogen, one chlorine, one nitrogen, and two oxygens. The substance also must contain *at least* one six-membered C_5N isolated ring system and have an Index Name which generates digram bit positions corresponding to those generated from the words CHLORO, CARBOXY, and PYRIDIN. The substance illustrated in Figure 6 is one example of a substance which satisfies the requirement of the screen search profile for the sample query.

In addition to individual query screens, the encoder may also specify screens for an entire batch of queries. In batch screening, the features common to all the queries in the batch are encoded. For example, if all the queries in the batch have a phenyl ring and a nitrogen atom, then this information can be encoded in a batch screen. The use of batch screens reduces screen search execution time when there is a significant number of query screens, since the query screens are applied only to those substances which satisfy the batch screens.

Text Search. For those substances that pass the screen search, text search is performed on the CA Index Names and related substance data such as molecular formulas and ring descriptions. The fundamental technique used in text searching is that of substructure search via nomenclature, in which nomenclature terms corresponding to structural features of the desired substructure are used as text search terms. The nomenclature screening in screen search determines if a given set of alphabetic digrams is likely to be present in a substance name, but cannot determine the correct digram sequence for the nomenclature terms. This is determined in text search by matching the nomenclature search terms on a character-by-character basis against the candidate CA Index Names.

The text search module in the search program allows for AND, OR, and NOT Boolean logic with two levels of nesting, e.g., A *AND* (B *OR* (C *AND* D)), where each letter represents one or more search terms. In addition, search terms may be

"CHLOROPYRIDINECARBOXYLIC ACIDS"

Logic	Terms
	CHLORO
AND	
	PYRIDIN
AND	
	CARBOXY

Figure 13. Text search profile.

weighted to give more importance to some terms than to others, or to permit negation of undesired terms.

Figure 13 shows a simple text search profile for the "chloropyridine carboxylic acid" query. The three search terms describe the basic structural units in the query. The asterisks surrounding the terms indicate truncations, which permit these terms to be identified when they are embedded in longer character strings. Thus, the term *PYRIDIN* would be found in both "pyridine" and "pyridinyl". The terms are connected by AND logic, which requires all three terms to be present in an Index Name for retrieval. AND logic is used since the terms may not be contiguous in a relevant CA Index Name because of intervening locant and nomenclature terms.

The substance illustrated in Figure 6 satisfies the specifications of the text profile for the sample query, since all of the required search terms are contained in the text string of its CA Index Name.

This sample profile is relatively generic since it permits truncated "stem" terms to be present anywhere in the CA Index Name; as a result, it is possible for some irrelevant substance-records to satisfy the profile requirements. A more precise, but somewhat more complicated, nomenclature text search profile can be constructed by using more exact terms and by searching for them in specific Index Name segments.[15] For example, searching for *CARBOXYLIC ACID* in the Heading Parent segment and *CARBOXY-* in the Substituent segment would prevent retrieval of an irrelevant environment for the "carboxy" term such as "2-carboxyphenyl". The text search module provides for specification of individual name segments. It is also possible to "ignore" certain typical irrelevant environments for search terms through the use of appropriate weighting logic. For example, searching for *PYRIDIN* in the Heading Parent segments but using weights to ignore *PYRIDINIUM* would prevent retrieval of one particular irrelevant environment for "pyridin". Developing these more precise nomenclature text search profiles requires a good understanding of the construction of CA Index Names. For example, a rationale for searching the Substituent segment using *CARBOXY-* would be preferred to one using *CARBOXY*, since in a relevant environment, "carboxy" will be followed by a hyphen to delimit it from the locant for the "chloro" substituent, e.g., in "(3-carboxy-4-chloro-2-pyridinyl)".

In searches where the substance class is to be correlated with nonsubstance or concept information in the textual descriptors, generic nomenclature profiles are typically used because the restrictiveness of the nonsubstance portion of the profile prevents the retrieval of most irrelevant substances. For example, if information relating to the metabolism of "chloropyridinecarboxylic acid" is desired, searching for *METAB* in textual descriptors would prevent the retrieval of the irrelevant substance, "3-Pyridinecarboxylic acid, 2-(4-chlorophenoxy)-" shown in Figure 14. The retrieval illustrated in Figure 14 is from a search of Volume 82 of the retrospective file with the sample profile (see Figure 13).

Link Search. In many cases, text search gives adequate search results, and no further processing is required. It is possible, however, for some irrelevant substance-records to pass text search because of "false coordination" of search terms. For example, a CA Index Name may contain all the required

"CHLOROPYRIDINECARBOXYLIC ACIDS"

REG= 4684-94-0 PMF= $C_6H_4ClNO_2$
NAM= 2-Pyridinecarboxylic acid, 6-chloro-
ELE= */C5N/*

metab. of
82: 68958S
⋮

REG= 51362-37-9 PMF= $C_{12}H_8ClNO_3$
NAM= 3-Pyridinecarboxylic acid, 2-(4-chlorophenyl)-
ELE= */C6/C5N/*

prepn. and cyclization of
82: 106184W

Figure 14. Volume 82 retrieval (partial).

"CHLOROPYRIDINECARBOXYLIC ACIDS"

Link	Terms/Logic
CONNECTIVITY	*CHLORO* *AND* *PYRIDIN*

Figure 15. Link search profile.

search terms, but the terms may not describe the proper arrangement of structural units. It is often possible to eliminate such irrelevant retrievals in the third level of search—link search.

Link search is a refinement technique rather than a necessary step in substructure searching via nomenclature. It has been defined as a separate search level, usually used only to reduce manual processing time. It is not needed for most searches since a high degree of precision can be achieved by using standard text search techniques. However, link search is generally employed in those relatively rare cases in which it is desirable to automatically eliminate a large number of irrelevant retrievals.

Link search is applied only to substances which satisfy the requirements of both screen search and text search. Since CA Index Nomenclature is highly structured, it is possible to deduce certain structural characteristics from the sequence of nomenclature terms, punctuation, and enclosing marks (i.e., parentheses and brackets) in a CA Index Name. For example, two nomenclature terms which are separated either by no enclosing marks or by matched pairs of enclosing marks usually describe connected structural units. Link search makes it possible to exploit some of these subtleties of CA Index Nomenclature by specifying attributes such as the number of occurrences of a search term, the order of search terms in a text string, and the number of characters which must or may separate two search terms.

The use of a substructure connectivity link is illustrated in the link search profile for the sample query shown in Figure 15. This link search profile will retrieve a relevant name such as "1,2-Benzenedicarboxylic acid, 4-(2-carboxy-6-chloro-4-pyridinyl)-", but will not retrieve an irrelevant name such as "1,2-Benzenedicarboxylic acid, 4-[2-carboxy-6-(2-chloroethyl)-4-pyridinyl]-". In the latter name, the "chloro" term is not at the same enclosing mark level as "pyridinyl", although "chloro" occurs before "pyridinyl", which is required for connectivity. The link search algorithm determines this fact from the presence of the unmatched right parenthesis between "chloro" and "pyridinyl".

SEARCH PERFORMANCE

Recall, precision, file screenout, and search execution times are common measures of performance for substructure search systems. As yet, we have not accumulated a large amount of data on these factors; our experience to date is summarized in the following sections.

Recall. Currently, the recall for substructure searches is

being verified only through spot-checking of the corresponding volumes of the CA Chemical Substance Index. However, the recall for a properly coded nomenclature substructure search should be complete relative to the results obtained from running the same search in a topologically based system. In our early experiments on nomenclature searching,[6] topological searches were run as controls to verify that nomenclature search techniques and search aids would result in complete retrieval. In addition, a more recent comparison of our retrieval results for a set of queries with the retrieval results obtained from a topologically based system, using a comparable database, found that the recall was the same.

Precision. Most of the retrospective searches CAS has conducted thus far have not included a precision-refining link search profile, and consequently there is, as yet, no firm data concerning the precision levels that are possible using this substructure search system. As noted earlier, profiles for test searches are not designed to ensure high precision, and the specifications for most of the retrospective searches conducted to date have been defined so narrowly that retrieval volumes have been small. High precision is a major concern only when retrieval volumes are large. The largest retrospective search CAS has run to date retrieved about 13000 substance-records. The results of this search required only a spot-check review since virtually all of the retrieved substance-records were relevant.

File Screenout. For most substructure search systems that operate on large retrospective files, a critical performance factor is the effectiveness of the search screens. An efficient set of screens minimizes the amount of searching that is needed in performing substructure searches. Searching can be a very time-consuming process, whether it is done using the combined text search/link search capability of the system described here or an iterative (atom-by-atom) search capability used in topologically based systems.

We have compiled statistics for the initial 46 substructure searches performed using the search system. Most of the queries were submitted by customers requesting a test search in our experimental search service. Eleven queries were obtained from an existing collection of questions at CAS which were used in our system test. Almost all of the queries were searched on at least one volume of the search files, i.e., 250000 substances.

The average screenout for the 46 queries was 98.6%, with a high of 100% and a low of 84.3%. These high and low figures were dropped in computing the 98.6% average. The low figure was for a query that consisted of three fragments: a phenyl group, a three-carbon chain, and one acyclic nitrogen. These three fragments apparently occur together in a significant number of substances.

One advantage of the screening mechanism in this system is that the screen specifications can be adjusted to some extent to improve screenout. If a test search with initial specifications results in low screenout, usually more precise nomenclature terms for the nomenclature screens can be used to improve the screenout. However, there is some tradeoff with screen search execution time when many screens are defined for a query.

In summary, we are obtaining very good screenout with a relatively simple and inexpensive set of screens. Of the three components of the substructure bit screen we have found that the nomenclature screen has the greatest relative screening power, followed by the ring system, and finally by the molecular formula screen. We are currently investigating other types of screens to improve our screenout even more and thereby reduce search execution time.

Ways of Reducing Search Execution Time. To serially process the total retrospective file of about 13 million records requires a minimum of 15–20 min of computer time on an IBM 370/168 computer. Consequently, we attempt to batch queries whenever possible to avoid multiple passes of the total file. We are currently investigating other ways to reduce computer time and other operational costs. Since the rate-determining step for most of our searches is the text search process, we are examining ways to speed up this step. In addition to the number of records to be searched, text search execution time is a function of the number of queries, the number and type of search terms, and the data elements to be searched. CAS is currently installing a state transition text search system which may be used in the substructure search system in the future. The most attractive feature of this text search algorithm is that the search time is essentially independent of the number of search terms. With this capability we expect to be able to conduct searches for single queries or for batches of queries with a large number of terms, using approximately the same amount of computer time.

Another possibility for reducing the cost of retrospective searches is to subdivide the search database by generic structural characteristics which occur commonly in substructure search queries. For example, the query specifications for many substructure search questions include either a carbocyclic or a heterocyclic ring system. Approximately 52% of the substances in the search database contain a carbocycle, while about 40% have at least one heterocyclic system (see Table I). If the substances in the database were divided into these two broad categories, storage requirements would increase significantly since many substances fall into both classes. However, overall search processing costs would decrease, since only one subfile would need to be searched for most queries.

The subfiles could be created by using the screen search capability to identify substances with the specified structural characteristics. At search time, the smallest subfile that contained substances having a desired characteristic would be used. For some queries, the total file would still need to be searched, but if the theory is correct, this would be necessary in only a small percentage of searches. We are currently examining a set of approximately 300 substructure search queries to determine typical substructure characteristics which might be used to subdivide the database.

CONCLUSION

This paper has described some of the important components of an experimental substructure search system at CAS based on CA Index Nomenclature. The system permits searches for all the substances and the associated descriptive information indexed for *Chemical Abstracts* since 1967. This large database makes it possible to correlate substructural data with descriptive data corresponding to chemical reactions, biological activities, and other applications.

Substructure searching using nomenclature represents only one of the approaches which CAS is exploring in its continuing search for more effective ways of disseminating chemical information. Undoubtedly, the techniques and the user aids which have been developed by CAS in the course of this work will become more important as CAS files containing CA Index Nomenclature become more widely available through the use of on-line systems.

REFERENCES AND NOTES

(1) R. E. O'Dette, "The CAS Data Base Concept", *J. Chem. Inf. Comput. Sci.*, **15**, 165–169 (1975).
(2) P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.*, **16**, 111–121 (1976).
(3) W. E. Cossum, G. M. Dyson, M. F. Lynch, and R. N. Wolfe, "Automation and Scientific Communication", H. P. Luhn, Ed., American Documentation Institute, Washington, D.C., 1963, pp 15–18.

(4) W. E. Cossum, M. L. Krakiwsky, and M. F. Lynch, "Advances in Automatic Chemical Substructure Searching Techniques", *J. Chem. Doc.*, **5**, 33–35 (1965).
(5) H. R. Schenk and F. Wegmuller, "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System", *J. Chem. Inf. Comput. Sci.*, **16**, 153–161 (1976).
(6) R. L. Wigington, "Machine Methods for Accessing Chemical Abstracts Service Information", Proceedings of IBM Scientific Computing Symposium on Computers in Chemistry, IBM Data Processing Division, White Plains, N.Y., 1969.
(7) W. Fisanick, L. D. Mitchell, J. A. Scott, and G. G. Vander Stouw, "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files", *J. Chem. Inf. Comput. Sci.*, **15**, 73–84 (1975).
(8) J. F. B. Rowland and M. A. Veal, "Structure-Text and Nomenclature-Text Searching for Chemical Information: An Experiment with the *Chemical Abstracts* Integrated Subject File and Registry System", *J. Chem. Inf. Comput. Sci.*, **17**, 81–89 (1977).
(9) W. Fisanick, "Substructure Searching of Files Derived from the CAS Chemical Registry System", Chemical Abstracts Service Open Forum, Philadelphia, Pa., April 1975.
(10) B. M. Vasta and M. L. Spann, "Chemical Searching Capabilities of CHEMLINE", presented to the Division of Chemical Information, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 31, 1976.
(11) P. F. Rusch and T. M. Crawford, "Improved Access to the Chemical Literature", presented to the Division of Chemical Literature, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 31, 1976.
(12) System Development Corporation, *Search Service News*, (5)1, Jan 1977.
(13) J. E. Blake, "Parent Compound Handbook—The Successor to the Ring Index", presented to the Division of Chemical Information, 173rd National Meeting of the American Chemical Society, New Orleans, La., March 22, 1977.
(14) M. D. Harrison, "Implementation of the Substring Test by Hashing", *Commun. ACM*, **14**, 777–779 (1971).
(15) N. Donaldson, W. N. Powell, R. J. Rowlett, Jr., R. W. White, and K. V. Yorka, "*Chemical Abstracts* Index Names for Chemical Substances in the Ninth Collective Period (1972–1976)", *J. Chem. Doc.*, **14**, 3–14 (1974); CA Volume 76 Index Guide, 1972.
(16) These techniques are now a major component of the BASIC Chemical Information System in Switzerland.[5]
(17) The full CAS Registry file is somewhat larger than this retrospective file, primarily because it contains substances which have been added by the registration of special substance files and which have not been indexed in *Chemical Abstracts* since 1967.

Noble Gas Chemistry and the Fluoride Literature—What Influences Research Directions?

DONALD T. HAWKINS* and WARREN E. FALCONER
Bell Laboratories, Murray Hill, New Jersey 07974

Received June 17, 1977

A discontinuity was observed in the growth rate of the literature on the free molecular structures of binary fluorides. It is suggested that this resulted from the landmark discovery by Bartlett in 1962 that stable compounds could be made containing noble gases. This discovery has had an influence not only in immediately related areas of chemistry, but also in areas several steps removed.

The question of what influences research directions, and indeed what stimulates research, is often asked. Only rarely is a cause and effect relationship identified. A surge of activity in one field or another is frequently observed. Sometimes this is the direct result of a new technique or a new concept; occasionally more subtle stimuli can be identified.

We have recently compiled a comprehensive bibliography on the molecular structures and force fields of binary fluorides.[1] As part of some statistical information for that bibliography, the number of papers cited in the bibliography for each year between 1955 and 1975 was determined. Figure 1, taken from the Introduction to the bibliography, shows these data.

From 1958 through 1961, the number of papers per year was constant at about 30 citations per year. There was a large increase to nearly 60 papers in 1962, and again to 85 papers in 1963. Since 1966 the yearly publication rate has been approximately steady near 100 papers per year. (The data for 1975 and 1976 are incomplete because of the time lag in the abstracting services used.)

We ascribe the discontinuity in the publication rate in 1962, and the increased number of publications since that time, to a surge of interest in and familiarity with reactive fluorides following the landmark discovery by Bartlett in 1961 that compounds of the noble gas xenon (previously thought to be chemically inert) could be synthesized.[2] The first order effect of this discovery was the well-known flood of research activity on noble gas chemistry itself which began in 1962 and is continuing actively to this day. A secondary effect could also be observed in the chemical literature; much more active synthesis and characterization in the whole spectrum of inorganic fluorides ensued in the immediate wake of Bartlett's synthesis of xenon hexafluoroplatinate. Free molecular

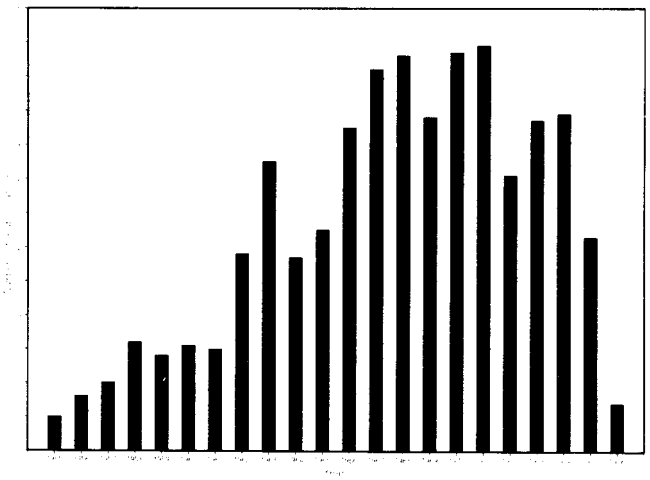


Figure 1. Papers cited per year.

structures, however, are at least one step further removed. Here we are dealing with physical measurements on a class of molecules which were generally not handled by chemists making structural measurements prior to the advent of noble gas chemistry in a real sense. The great interest in characterizing the structures and bonding of the noble gas fluorides, and the experimental and theoretical experience gained with these molecules, apparently spilled over to encompass a much wider selection of binary fluorides and related molecules.

It is rare that one can point to a single event which influences a broad area of chemistry. More often, one observes slow growth in a field, a major discovery, and then further steady growth as the effects of the discovery spread.[3] In this