Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files†

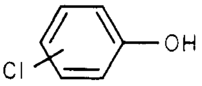
W. FISANICK,* L. D. MITCHELL, J. A. SCOTT, and G. G. VANDER STOUW

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received March 6, 1975

The increasing availability of computer-readable files of chemical nomenclature and of programs for text searching has led to the development of methods for performing substructure searches in which CA nomenclature terms are used as search terms. Substructure searches on CA Index nomenclature can often result in very high recall relative to topological searches, as is shown by experimental results achieved on a variety of searches. Many data bases which contain CA Index nomenclature also contain nonsubstance data. Thus, searching of substance and nonsubstance data can often be done within a single search of a file with both high recall and relevancy. Profile construction aids prepared by CAS make it possible for persons without sophisticated nomenclature backgrounds to construct nomenclature profiles for many questions.

In the Chemical Substance Index to *Chemical Abstracts* (CA), a chemical substance is described by its CA Index Name, which is prepared by application of a rigorous set of nomenclature rules.[3] The great majority of CA Index Names are constructed according to systematic nomenclature rules; that is, they are constructed from nomenclature terms which correspond to fragments of a chemical structure. A vocabulary of a few hundred such terms, representing rings, chains, and functional groups, is used to form names for millions of chemical structures. Consequently it is possible to use Chemical Substance Index nomenclature as a data base for substructure searches in which CA Index Names are examined to find nomenclature terms or sets of terms which correspond to a desired substructure. Performing such manual searches on a printed file such as the Chemical Substance Index is frequently difficult because the desired substances may be placed at several different headings, some of which are not at all obvious. If, for example, one is searching for the chlorophenol substructure



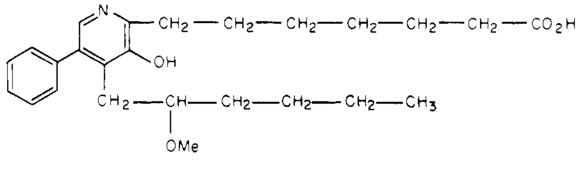
some answers would be found by examining names indexed under "Phenol" to find those containing "chloro". However, the question is equally well answered by such names as "1-Anthracenesulfonamide, 4-(3-chloro-5-hydroxyphenyl)-" and "9*H*-Xanthene-3-carboxylic acid, 8-methoxy-1-methyl-9-oxo-, (3-chloro-5-hydroxyphenoxy)methyl ester". There is no *a priori* means for identifying such answers; finding them in a manual search would essentially require scanning all names in the index.

In recent years several computer-readable files containing CA Index Names have become available. Among these are Chemical-Biological Activities, Polymer Science & Technology, Energy, Materials, Food and Agricultural Chemistry, Ecology and Environment, Chemical Abstracts Integrated Subject File, and Chemical Abstracts Subject Index Alert from Chemical Abstracts Service (CAS), as well as the CHEMLINE on-line retrieval file set that is available in the TOXLINE service of the National Library of Medicine.[8,9] The CA Index Names and related data on these files can be searched with text search (character matching) programs. With such programs, substructure searches can be performed using text search profiles in which the search terms are chemical nomenclature terms that describe a desired substructure. The answers retrieved will be names that contain the desired terms and therefore should describe structures that contain the desired substructure. In this way one can retrieve names that are not readily accessible through a manual search of a printed alphabetic listing. For example, the names shown for the manual "chlorophenol" search could all be retrieved by a search profile containing search terms such as Chloro, Hydroxy, Phenyl, Phenol, and Phenoxy and using appropriate Boolean logic and term truncation.

In this paper we describe an investigation of nomenclature-based substructure searching using techniques and search aids developed by CAS. The procedures and search aids discussed here are designed for searching files containing CA Index Names for the Ninth Collective Index (9CI) period (Volumes 76–85, 1972–1976)[1,5] and are an extension and modification of procedures and search aids developed for the Eighth Collective Index (8CI) period (Volumes 66–75, 1967–1971).[2,6]

NOMENCLATURE TERMS AND SUBSTRUCTURES

Systematic Index Names. Most chemical substances indexed from CA and named according to 9CI rules receive highly systematic names which are constructed from terms that correspond to structural fragments such as rings, chains, and functional groups. For example, the name "2-Pyridineheptanoic acid, 3-hydroxy-4-(2-methoxyhexyl)-5-phenyl-" for the structure


N
CH2 — CH2 — CH2 — CH2 — CH2 — CH2 — CO2H
OH
CH2 — CH — CH2 — CH2 — CH2 — CH3
OMe

is formed from the terms "pyridine" and "phenyl" which represent rings, "heptan" and "meth" which describe carbon chains, and the functional group terms "oic acid" and "hydroxy".

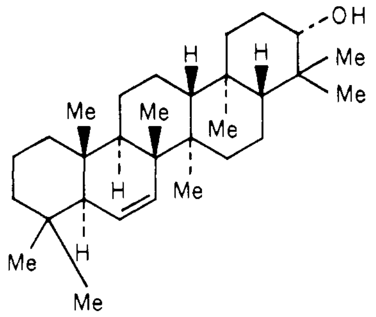
In systematic CA Index Names, a specific structural unit is usually described by different terms depending on whether it is named in the Heading Parent segment of a name or in the Substituent segment. This fact is a consequence of the hierarchy of compound classes which is used in the nomenclature rules.[3] Application of this hierarchy,

† Presented, in part, at the 167th National Meeting of the American Chemical Society, Los Angeles, Calif., April 2, 1974.

* To whom correspondence should be addressed.

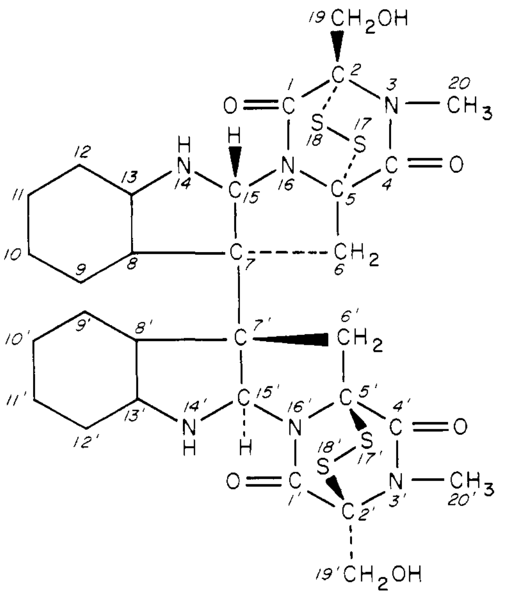
or order of precedence, identifies one functional group in a substance as the principal functional group, and this group is named in the Heading Parent segment of the name, usually along with a ring or chain unit attached directly to it. Other functional groups and skeletal fragments are named in the Substituent segment. Examples in Figure 1 illustrate the use of different terms for a given structural unit. In the first example, the OH group is the principal functional group and is described in the Heading Parent by the suffix "ol". In the second example, the OH is subordinated to the higher-ranking carboxy group and therefore is denoted by "hydroxy" in the Substituent segment. Note that the heterocyclic unit terms have the same root in both instances. The use of multiple terms to describe a unit needs to be taken into account when developing a nomenclature substructure search profile.

Semisystematic Index Names. Some CA Index Names can be described as semisystematic. Such names contain terms which do not correspond directly to rings, chains, or functional groups but rather to combinations of these structures which could be given fully systematic names. For example, in the name "Gammacer-15-en-3-ol,(3β)-" for the structure



the nonsystematic term "gammacer" describes both the ring and the attached methyl groups while the systematic suffix "ol" describes the attached hydroxy unit. These "semisystematic" names usually involve stereoparents such as "Gammacerane". (A stereoparent is an index heading parent, the name of which implies specific stereochemical information.)

Nonsystematic Index Names. In a few cases, CA Index nomenclature uses nonsystematic names, i.e., names which do not at all indicate the structure being described. For example, the common name "Chetocin" is used to describe the specific stereoisomer



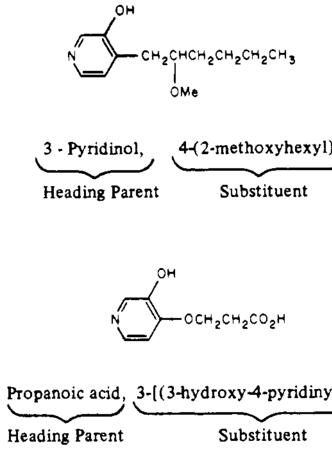


Figure 1. In a CA Index Name, the choice of a term used to describe a given structural unit depends on whether the unit is named in the Heading Parent or the Substituent segment of the name.

Since not all CA Index Names are fully systematic, a nomenclature search profile must take into account not only the name terms used in fully systematic nomenclature but also, if high recall is desired, terms used in semisystematic or nonsystematic names. The profile should also recognize the possibility of "false drops" where one nomenclature term is fully embedded within another, as, for example, the term "ethyl" in "methyl", or "thiazole" in "isothiazole" and "oxathiazole". The construction of a search profile for nomenclature search is thus potentially a complex task.

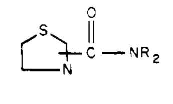
SEARCH AIDS

To assist the searcher in identifying appropriate nomenclature search terms, in determining possible undesired terms, and in using this information to establish an effective text search profile, CAS has developed several experimental aids for 9CI nomenclature searching.[1,5] (The CA Index Guide and Chemical Substance Index are of course also useful aids in deriving search terms.) Some of these search aids were produced by computer program from 9CI data files; others were prepared manually. Brief descriptions of the major nomenclature search aids are given below. In these descriptions, the use of the aids is described for the sample substructure search query shown in Figure 2.

The manual "Substructure Searching of Computer-Readable CAS Chemical Nomenclature Files"[5] describes search strategy and profile construction techniques for searching nomenclature files. This manual also contains a detailed discussion of the use of other specialized user aids.

The "Keyword-Out-Of-Context (KWOC) Index of Ring Systems" is a computer produced list derived from the CAS Index of Ring Systems data base. This user aid provides access to both systematic and nonsystematic nomenclature terms for ring systems via a component ring formula, i.e., a formula which gives the number of atoms of each element present in an individual ring. Hydrogen is not included in this formula. The thiazole ring system from the sample search question (Figure 2) has the ring formula C_3NS. Examining the KWOC Index at the C_3NS formula (see Figure 3) leads to the names of isolated ring systems having that formula and, also, to the names of multiple ring systems which contain a C_3NS ring as a basic component.

The KWOC Index is particularly useful in deriving search terms for substructure searches for ring systems in which fusion to the desired ring system is allowed; i.e., the desired ring may be embedded in a larger ring system. For example, in the sample question (Figure 2) the ring system 3-thia-1-azabicyclo[3.1.0]hexane



- R = H, CH, CH_2, or CH_3
- Ring fusion allowed
- Any degree of unsaturation in the ring is allowed
- Complete recall desired

Figure 2. Sample substructure search via nomenclature query.

→ C_3NS	C_3NS	Isothiazole	Isolated ring systems
	C_3NS	Thiazole	
	⋮	⋮	
	CNS-C_3NS	1,4-Dithia-2,7-diaza= spiro[2.4]heptane	Multiple ring systems
	C_2N-C_3NS	3-Thia-1-azabicyclo= [3.1.0]hexane	
	⋮	⋮	
	CN_4-C_3NS	Thiazolo[3,2-d]tetrazole	
	⋮	⋮	

Figure 3. Keyword-Out-Of-Context (KWOC) Index of Ring Systems.

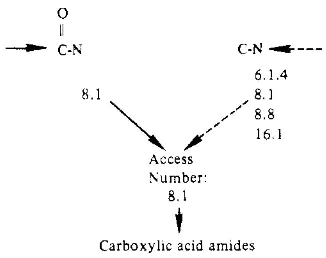


Figure 4. Functional Group Guide (Part 1).

8.1	CARBOXYLIC ACID AMIDES
PF	SF
•CARBOXAMIDE	•AMINO• AND •CARBONYL• •OXO-• AND •AZA• ⋮

Figure 5. Functional Group Guide (Part 2).

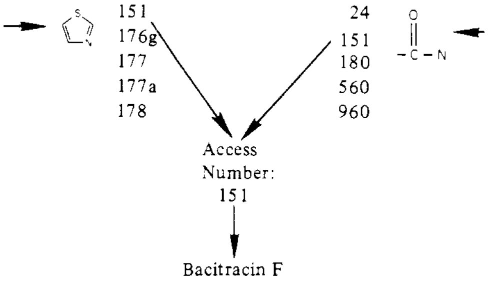


Figure 6. Nonsystematic Name Guide.

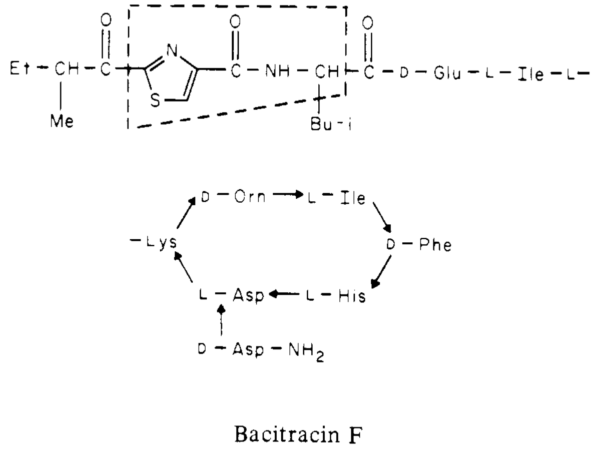
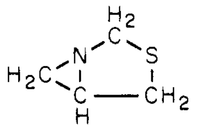


Figure 7. Nonsystematic name and structure for substance containing thiazolecarboxamide substructure.

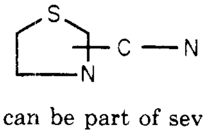


is relevant since it contains the thiazole system and therefore should be accounted for in the search profile if total recall is desired.

The "Functional Group Guide" is a manually derived search aid based on CA Index nomenclature rules. This guide provides access to terms which are used in systematic nomenclature to describe functional groups. Only functional groups which appear in the Order of Precedence of Compound Classes used in the CA Index nomenclature rules are included. In the guide, nomenclature terms for functional groups are located via a structural diagram. The structural diagram leads to a list of identification numbers (see Figure 4) which correspond to those groups which are identical with *or contain* the group of interest. Figure 4, for example, shows that the amide group leads to the number 8.1. The same number is also listed for the more generic C–N fragment.

The identification numbers lead to the nomenclature terms which appear in another section of the guide. There the number 8.1 leads to the Carboxylic Acid Amides section which contains two lists of terms (see Figure 5). The PF list gives the nomenclature fragments which describe the carboxamide group when it is the principal functional group; the SF list gives terms used when the group is a subordinate function.

The primary importance of the Functional Group Guide is that it helps the user to derive nomenclature terms which describe subordinate functional groups; such terms are otherwise difficult to determine, even for a nomenclature specialist. For example, if the substructure being sought is



the acyclic portion can be part of several other functional groups besides carboxamide such as amine, azide, and hydrazide. The Functional Group Guide provides access to names for these groups via the C–N fragment.

The "Nonsystematic Name Guide" is a manually derived search aid based on the CA Index Guide. This guide is accessed via structural diagrams which correspond to ring, chain, or functional group structural units. Each structural unit is associated with a list of identifiers which lead to another section of the guide. That section contains nonsystematic names which describe substances which contain that unit. Only nonsystematic names which are used as Heading Parents in the CA Chemical Substance Index are included. As shown in Figure 6, the lists which correspond to the thiazole ring and to the amide group both contain 151 which points to the name "Bacitracin F" (see Figure 7).

The Nonsystematic Name Guide is a valuable search aid for queries in which the search goal is complete recall since a few nonsystematic terms are generally needed to provide complete identification of a substructure via nomenclature.

The "Key-Letter-In-Context (KLIC) Index" is a computer-produced list based on Volume 77 of the CA Chemical Substance Index. This KLIC list consists of words, i.e., strings of contiguous alphabetics, used in CA Index nomenclature. Each alphabetic string is rotated so that a word in the KLIC may be located via each letter in the string except the last one. The KLIC Index helps the user determine what terms are likely to be retrieved by a given truncated term. For our sample search (Figure 2), the KLIC

.	.
Thiazocine	104
.	.
Benzo Thiazolamine	85
Thiazole	325
.	.
Iso Thiazole	53
.	.
Benzo Thiazole	377
.	.
Methoxybenzo Thiazolium	2
.	.
Thiazolo	340
.	.

Figure 8. Key-Letter-In-Context (KLIC) Index.

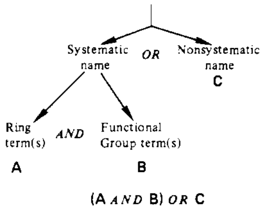


Figure 9. Basic logic of nomenclature search profile.

Index (see Figure 8) shows that Thiazol with left and right truncation will retrieve a number of relevant terms but also the irrelevant "isothiazole"; further right truncation would probably lead to additional irrelevant retrievals such as "Thiazocine". The KLIC Index also gives the frequency of occurrence of each term in the CA Chemical Substance Index to Volume 77, thus providing a guide to estimating the volume of retrieval which may be expected from a specific search profile.

SEARCH PROCEDURE

The procedure of performing a nomenclature substructure search can be described as five basic steps, in addition to the actual running of the text search program:

1. Development of basic search strategy or logic
2. Nomenclature term derivation
3. Search term derivation
4. Profile construction
5. Performance optimization

Development of the Basic Search Strategy. The basic approach used in preparing a nomenclature search profile depends on the type of search performance desired. If complete recall is not needed, it is usually adequate to use a small number of systematic nomenclature terms in the profile. However, if complete recall is desired, it is necessary to anticipate all the ways in which a specific structural unit may be named. Preparing a profile for high recall will generally require the use of some of the profile construction aids which were mentioned above.

The structure in the sample query (Figure 2) has two basic components: the ring and the attached functional group $-C(=O)NR_2$. A systematic name for a substance which contains this structure will contain a term or terms which describe the ring and a term or terms which describe the functional group. Since both the ring and the functional group must be present, the search question will include terms describing the ring and terms describing the functional group. These two sets of terms will be connected by AND logic, i.e., a term or term combination from each set must be present in a substance name to cause retrieval of that name. If the ring terms are symbolized by A and the functional group terms by B, the systematic part of the search profile is (A AND B).

Ring term(s)		Functional group term(s)
A	AND	B

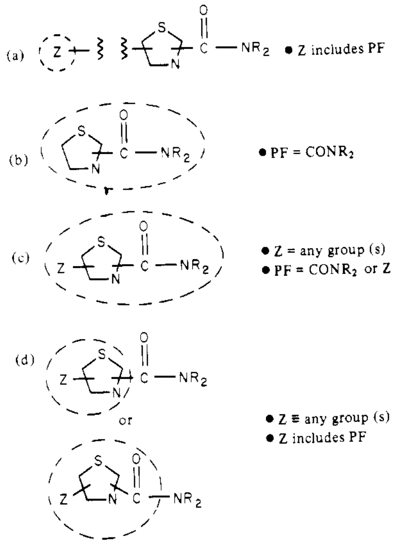


Figure 10. Possible locations of desired nomenclature terms within data elements of CA Index Name. (In each case, the structure named in the Heading Parent is circled by a dashed line.)

An AND logic connection is needed since the ring terms and the functional group terms in a given name will not necessarily be contiguous. For example, in the name "Benzoic acid, 4-[4-(aminocarbonyl)-5-chloro-2-thiazolyl]-", the ring term "thiazolyl" and the functional group term "aminocarbonyl" are separated because of the alphabetical ordering of the radical prefixes.

Although almost all substances that contain the desired substructure will receive systematic names, the substructure may also occur in a substance described by a nonsystematic name. Nomenclature terms derived from nonsystematic names should also be included in the search profile to ensure complete recall. These nonsystematic terms would be connected to the systematic terms by OR logic; i.e., the presence of either the systematic term combination or a nonsystematic term would cause retrieval of a substance name. If the nonsystematic names are symbolized by C, then the logic of the search profile becomes (A AND B) OR C as shown in the tree structure in Figure 9.

If the specific segments, or data elements, of a CA Index Name are to be searched individually, one needs to consider four possible ways in which the desired nomenclature terms can be located within these data elements. The first case is that in which both the ring and the functional group are described in the Substituent or Name Modification segment and some different portion of the substance structure is described by the Heading Parent. This case can be represented as shown in Figure 10a (PF means principal function in Figure 10). Two examples of this type of name are:

$$\text{Benzoic acid, 4-[4-}\underbrace{\text{(aminocarbonyl)}}_{\text{functional group}}\text{-2-}\underbrace{\text{thiazolyl}}_{\text{ring}}\text{]-}$$

$$\text{Benzoic acid, 4-}\underbrace{\text{[(methylamino)carbonyl]}}_{\text{functional group}}\text{-2-}\underbrace{\text{thiazolyl}}_{\text{ring}}\text{ ester}$$

The second case (Figure 10b) is that where the structure named by the Heading Parent is exactly the desired substructure. Such names can be retrieved by search terms such as "Thiazole" to represent the ring and "Carboxamide" to represent the functional group. An example of a Heading Parent which will satisfy the example question is "2-Thiazolecarboxamide".

In the third case (Figure 10c) the Heading Parent names a structure which contains, but is larger than, the desired substructure. For example, the name "3-Thia-1-azabicyclo[3.1.0]hexane-2-carboxamide" describes a structure which contains the thiazole ring of the sample question embedded in a larger ring system. "Bacitracin F" (Figure 7) is an example of a nonsystematic name for a substance which contains the "thiazolecarboxamide" substructure.

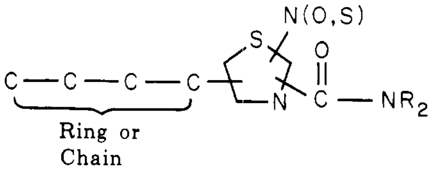
The fourth and most complex case (Figure 10d) is where the desired substructure is partly described in the Heading Parent and partly in other data elements. The following two names are examples of this case:

$$\text{2-Benzo}\underbrace{\text{thiazole}}_{\text{ring}}\text{acetic acid, 4-(}\underbrace{\text{aminocarbonyl}}_{\text{functional group}}\text{)-}$$

$$\text{4-}\underbrace{\text{Thia-1-aza}}_{\text{ring}}\text{bicyclo[3. 2. 0]heptane-2-carboxylic acid, 6-[(dimethyl}\underbrace{\text{amino)carbonyl}}_{\text{functional group}}\text{]-}$$

If complete recall is an objective of a search, all four possibilities shown in Figure 10 must be considered.

Another important strategy consideration is the time needed to encode the profile. Certain structural units require more time to encode then others. For example, consider a search for

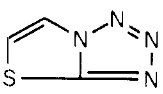


● No ring fusion allowed

Here the N(O,S) and C–C–C–C units are very generic and difficult to encode. For the N(O,S) unit, search terms for amines, amides, ethers, esters, sulfides, nitrogen heterocycles, etc., would be needed. For the C–C–C–C unit, search terms for acyclic hydrocarbon chains, e.g., butanes, pentanes, hexanes, and heptanes, would be needed along with terms for a large number of ring systems. To avoid time-consuming search term derivation for these units, the user might, instead, frame the question on a more generic level and ignore the N(O,S) and/or C–C–C–C unit in the encoding. This would greatly simplify profile construction and would not adversely affect recall. In general, search terms for two or three units of a multiunit substructure are sufficient to retrieve the desired names along with a tolerable number of irrelevant names.

Nomenclature Term Derivation. Once the basic strategy for a search is determined, the next step is to choose the nomenclature terms that correspond to the desired structural units. In the sample question (Figure 2), terms which describe the ring system, terms which describe the functional group, and nonsystematic name terms which describe substances containing both groups must be determined. Some examples of these terms and their relation to the logic expression are shown in Figure 11. These terms were derived with the help of the specialized user aids described above.

Ring Terms. In the sample question the ring system presents a relatively complex situation. If the ring occurs as an isolated unit, it will be described by terms such as "thiazole", "thiazolyl", and "thiazolidine". A ring may also occur embedded within larger units, such as the fused ring system for our example (thiazolo[3,2-*d*]tetrazole).



The ring or "A" set of terms (see Figure 11) will therefore be made up of several different ring terms connected by OR logic; each term in the set will describe some ring which answers the question.

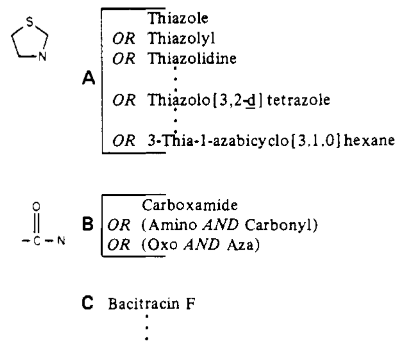


Figure 11. Nomenclature terms which describe the sample query.

Systematic names for those rings which contain a desired ring, either as an isolated ring or an embedded unit, are usually obtained from the KWOC Inverted Index of Ring Systems which is discussed above.

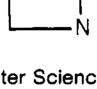
Functional Group Terms. In determining the nomenclature terms which will be used to describe a functional group, one must consider not only the case in which that group is the principal function and is named in the Heading Parent but also the cases where the group is subordinated to some other functional group. In the sample question, the desired group may be named by the term "carboxamide" when it is the principal function; if it is a subordinated function the terms "amino" and "carbonyl" or the terms "oxo" and "aza" will be used. Terms in the "B" group (see Figure 11) describe alternative ways of naming the functional group and are connected by OR logic; some of these terms require AND logic internally.

Systematic functional group terms are usually obtained from the Functional Group Guide (see Figures 4 and 5).

Nonsystematic Terms. "Bacitracin F" (see Figure 7) is an example of a nonsystematic CA Index Name that contains the thiazolecarboxamide substructure. When complete recall is required these "C" terms (see Figure 11) must be included. Nonsystematic names for substances that contain a desired structural unit are usually obtained from the Nonsystematic Name Guide.

Search Term Derivation. Although nomenclature terms may be used directly in the search profile, a search can usually be made more effective and efficient by modifying these nomenclature terms. In particular, it is often possible to use the technique of truncation, i.e., the extracting from several name terms of a common character string which can be used as a search term to retrieve the nomenclature terms. For example, for the sample question, the search term *THIAZOL* (the asterisks indicate truncation) will retrieve "thiazole", "thiazolidine", "thiazolyl", and "thiazolo".

The capability to use, or somehow to emulate, both left- and right-hand truncation is very important for effective nomenclature searching because nomenclature terms are frequently embedded in larger character strings. For example, as seen in Figure 8, "amine" is embedded in "benzothiazolamine". Thus, in a search for the amine functional group left-hand truncation should be used; i.e., search terms such as *AMINE or *AMINO* should be used. However, search terms with left truncation (or both left and right truncation) can sometimes retrieve larger, undesired name terms. For example, the use of *THIAZOL* for the structure



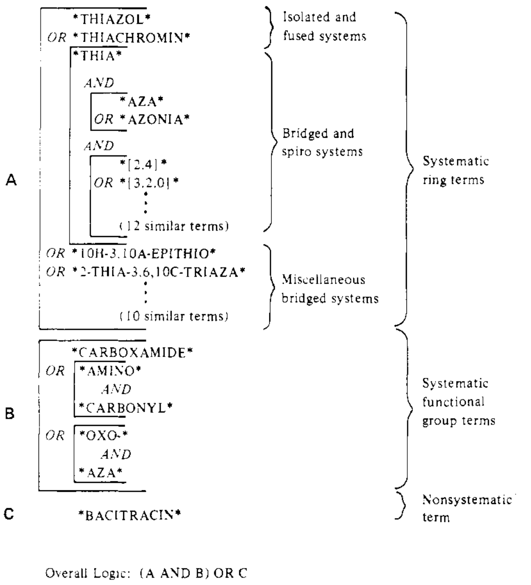


Figure 12. Portion of a search profile for the sample query.

can retrieve the undesired "Isothiazole", *ETHYL* for the C–C unit can retrieve the undesired "Methyl", and *ONE* for the C–C(=O)–C unit can retrieve the undesired "Sulfone".

The KLIC Index can guide effective use of truncation by identifying terms to be ignored or negated on the profile and helping the user estimate the volume of retrieval to be expected from a particular search term.

Profile Construction. The way in which a profile is constructed for a nomenclature substructure search depends on the search emphasis, i.e., whether high recall is a goal, and the specific search system being used (the logic capabilities available, whether individual search fields can be addressed, whether the search is to be done in batch or interactive mode), and on the organization of the file. Figure 12 shows a portion of a profile developed for the sample question (Figure 2). In writing this profile the following assumptions were made about the search, search system, and file organization:

(a) High recall emphasis
(b) Basic Boolean logic capability with the nested expressions allowed
(c) Four modes of truncation available (left, right, both left and right, and none)
(d) Batch mode, serial search of a file in which an index name is contained in a single search field

In the profile, all characters are shown as capitals; superscripts and subscripts are shown as on-line characters. The brackets on the left indicate the level of the Boolean operation, and the asterisks indicate truncation.

The first two terms in the profile are intended to retrieve the names of isolated and fused thiazole systems. The next set of terms is designed to retrieve most of the bridged and spiro ring systems. This set combines prefixes which describe nitrogen and sulfur atoms with numerics which describe bridges in relevant ring systems. For these systems, common character strings in the ring system names are used instead of a specific term for *each* of the ring system names. This compacts the profile since a large number of individual terms would otherwise be required. For example, among the two-ring systems alone there are over 40 ring systems with an embedded "thiazole" substructure. Since both the two heteroatoms (nitrogen and sulfur) and ring size specifications are required in relevant bridged and spiro systems, the corresponding sets of terms are connected by AND logic. The remaining ring terms are intended to retrieve other bridged systems. Here, a contiguous character subset from ring system names is used. Only as much of the name as is necessary to retrieve it with good precision is used, for example, as in the use of *10H-3,10A-EPITHIO* to retrieve substance names containing the ring system name, "10*H*-3,10a-Epithiopyrazino[1,2*a*]indole".

To ensure high recall, the profile shown in Figure 12 has been constructed on a relatively generic level, and consequently some irrelevant answers (false drops) can be anticipated.

If the individual segments of a name are contained in separate search fields or are delimited by special, nonnomenclature characters, it is usually possible to write a more precise profile, although the search logic becomes somewhat more involved. For the sample query (Figure 2), the data elements specified for each of the four cases discussed above (see Figure 10) would be searched and OR logic would be applied among the cases. Data element specification improves precision because it prevents some of the irrelevant retrieval that can arise when a single name field is searched with two or more nomenclature terms connected by AND logic. For example, the use of *THIAZOL* and *AMINO* and *CARBONYL* without data element specification could retrieve irrelevant names such as "2-Thiazolecarboxylic acid, 4-[(methoxycarbonyl)methoxy]-, compound with 4-aminobenzamide" with incorrect terms and name segment correlation.

If complete recall is not required, a simple profile can generally be used with few valid answers lost. The profile can be simplified by including only the terms by which the substructure is "most likely" to be named and by searching the data elements in which these terms are "most likely" to be found.

Since nonsystematic names for substances are relatively rare in the CA Index files compared to systematic ones, nonsystematic terms for a substructure can usually be ignored when complete retrieval is not necessary. Also, certain systematic terms occur more frequently than others. For example, in the profile for the sample question (Figure 12), the replacement ("a") nomenclature functional group terms *AZA* and *OXO-* can be expected to retrieve only a few, if any, carboxamides, while the other "B" terms are derived from substitutive nomenclature which is by far the predominant type of systematic nomenclature.

A precise profile for a substructure query can usually be developed by using, as search terms, character strings which express the "connectivity" among the structural units in the query. Some examples of precise terms for the sample query are *2-THIAZOLECARBOXAMIDE*, *4-THIA-1-AZABICYCLO[3.2.0]HEPTANECARBOXAMIDE*, and *AMINO)CARBONYL)-2-THIAZOL*. However, for queries with more than one structural unit it is usually not possible to identify all of these continuous character strings, primarily because of the alphabetical ordering of radical prefixes used in index nomenclature. Consequently, relevant substance names could be missed.

Performance Optimization. When complete recall is the objective, a good search technique is to do an initial search using a relatively generic profile such as the one illustrated in Figure 12. After examining the results, one can adjust the profile to improve relevancy. In some cases, the results obtained with an initial, generic profile will be sufficiently satisfactory so that the extra effort required to refine the profile will not be needed.

There are several ways in which a profile can be adjusted to improve relevancy without adversely affecting recall. One is to use search terms with more characters or to use a larger number of terms for a given structural unit. Another way is to encode all the structural units in the query if some of them were not included in the initial coding. If the system has expanded logic capabilities, these can often provide additional context for the search terms and conse-

Table I. Characteristics of Substructure Queries Used in Experiment

Characteristic	No. of queries
Acyclic only	8
Cyclic only	9
Monocyclic systems	
Phenyl	8
Heterocyclic	7
Fused systems	
Two rings	10
Three rings	4
Four or more rings	3

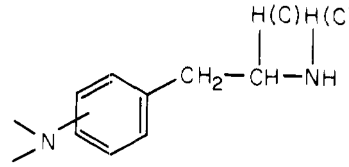
Table II. Average Atom Counts

All atoms	10.6	Phosphorus	0.05
Carbon	8.0	Sulfur	0.15
Nitrogen	1.2	Halogen	0.08
Oxygen	1.1	Metal	0.03

quently improve relevancy. Some system features that facilitate nomenclature substructure searching are discussed in the section titled Search System Features.

Another way to increase relevancy is to negate specific substances that can occur as "hits" but are irrelevant to the search. This is most simply done by negating, via the application of NOT logic, specific CAS Registry Numbers if these numbers are available in a searchable field. It is also possible, though less convenient, to negate a substance by negating its unique CA Index Name. The entire name should be negated and not just some of its name terms since partial negation might inadvertently prevent retrieval of some relevant names.

The CA Volume or Collective Chemical Substance Indexes are useful aids in identifying likely irrelevant answers that might be candidates for negation. For example, consider a query for



● Trivalent, singly bonded N

● No fusion

and a strategy using search terms for only the ring and the $-CH_2-\dot{C}H-\dot{N}H$ group. Such a strategy avoids the time-consuming derivation of name terms for the $-N<$ unit. However, for the case where the ring and the $-CH_2-\dot{C}H-\dot{N}H$ group are described in the Heading Parent, all the names for substituted phenethylamines that did not contain a substituent with a nitrogen atom attached to the phenyl ring would be potential irrelevant answers. An examination of the "Benzeneethanamine" heading in a CA Chemical Substance Index (Volume 78 was used for this illustration) indicates that there is one irrelevant substituted "Benzeneethanamine" name that has a large number of entries (i.e., has a high probabilfty of occurring): "Benzeneethanamine, α-methyls." If the Chemical-Biological Activities data file is being searched, then "Benzeneethanamine, α-methyl-" is even more likely to occur, since most of the text modifications at this heading are concerned with biological activity.

BASIC SEARCH EXPERIMENT

Design. The basic techniques and strategies used at CAS for nomenclature substructure searching of both 8CI and 9CI files were developed in an initial study performed in 1971 that involved serial batch searching of two files of 8CI names. One test file contained 2531 substances corresponding to one in every eight substances on the Common Data Base (CDB). (The CDB, a subset of the CAS Chemical Registry, contains approximately 22,000 substances, primarily drugs, dyes, pesticides and other commercially used substances.) On this file individual name segments are not separately addressable. The other test file contained 11,108 substance names corresponding to a subset of the substances represented in the Volume 71 CA Chemical Substance Index. This subset was composed of those Volume 71 Substance Index Names which also appear in the CDB or Chemical-Biological Activities. In this file, each segment of a name was addressable; i.e., name segments were contained in separate data elements.

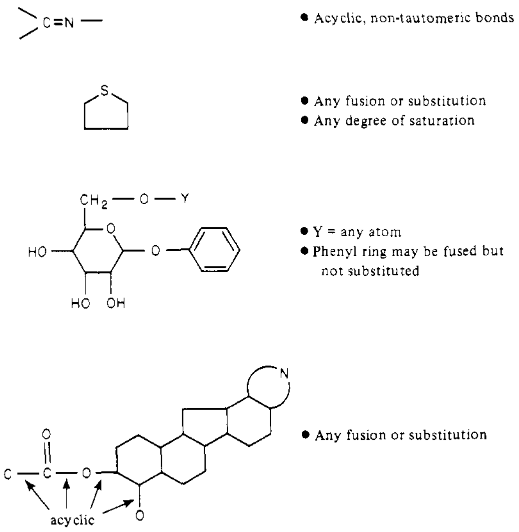

C=N —
● Acyclic, non-tautomeric bonds
S
● Any fusion or substitution
● Any degree of saturation
CH2 — O — Y
O
HO
O
HO OH
● Y = any atom
● Phenyl ring may be fused but not substituted
N
O
C — C — O
acyclic
O
● Any fusion or substitution

Figure 13. Query examples from experiment.

The CAS OS/360 Text Search System[7] was used to search the CDB subfile. These text search programs permit four modes of truncations and cumulative weighting but not nested Boolean expressions.

To search the Volume 71 Index subfile, the larger of the test files, a set of experimental programs which performed the function of a text search system was used. These programs allowed four modes of truncation (i.e., left, right, both left and right, and none) and nested Boolean expressions but did not permit cumulative weighting.

A set of 37 substructure search questions was run on each test file. These queries covered a variety of substructures ranging from portions of functional groups to substructures which include a steroid nucleus. Most of the queries were obtained from substructure search experiments involving topological substructure searching on connection table representations of substances with the CAS Substructure Search Programs.[4,10] Eight queries, which were also framed for topological searching, were obtained from the computer center at the University of Georgia, while another ten queries were generated at CAS especially for this study to provide additional variety to the total set. Some examples from this substructure query set are shown in Figure 13; some characteristics of the substructures are listed in Tables I and II.

The nomenclature search profiles for each query were constructed to provide complete recall; i.e., supporting user aids were used to develop a complete set of nomenclature/search terms for the structural units selected for encoding. To ensure high recall and to simplify profile construction the questions were framed on a highly generic level such as was illustrated in the profile for the sample search (Figure 12). Queries were grouped and run in four batches with approximately the same number of queries in each batch. Each substructure search was performed only once with the

Table III. Nomenclature and Fragment Search Results[a]

Nomenclature search	CDB subfile	Vol 71 index subfile
Total retrievals	1027	3920
Overall relevancy	64.5%	64.0%
Relevancy	67.0%	64.5%
Overall recall	97.9%	99.1%
Recall	97.5%	97.9%
File screenout	98.9%	99.0%
Fragment search		
Total retrievals	2047	6974
Overall relevancy	33.5%	37.1%
Relevancy	56.0%	48.8%
File screenout	97.8%	98.3%

[a] *Relevancy* is obtained by averaging the mean relevancies of the individual queries. *Overall relevancy* is the total number of relevant items retrieved for all queries divided by the total number of items retrieved.

Table IV. Relevancy Failures

Type of failure	CDB subfile	Vol 71 index subfile
1. Required structural unit not present in retrieved substance	48.7%	56.2%
2. Retrieved substance has unwanted substitution on required structural unit	17.2%	21.7%
3. Retrieved substance has required structural units present but incorrectly attached	20.3%	10.8%

exception of reruns to correct obvious coding errors. Queries were not reformulated and repeated to optimize the results.

Most of the encoding strategy for the profiles was developed by a staff member with about two years experience in naming substances according to CA Index nomenclature practices. However, the actual encoding was performed by a staff member with only a limited background in index nomenclature.

To provide a control for this experiment, searches for the same substructures were performed using the CAS Substructure Search Programs on search files containing connection table representations of the same substances that were in the two nomenclature files. CAS Substructure Search Programs allow two levels of searching: (1) fragment search, searching a bit indicator file containing indications of various structural fragments in the encoded substances, and (2) iterative search, searching, atom-by-atom, a connection table file. For most substructure searches, the fragment level search is used as a screen for the more time-consuming, but precise, iterative search. The recall of the fragment and iterative searches and the relevancy of the iterative searches were assumed to represent 100% as a standard for the nomenclature searches.

Analysis of Results. Table III shows the results of the nomenclature and fragment searches for the 37 queries. Complete recall was obtained for 33 out of the 37 queries in searching the CDB subfile and for 29 out of the 37 queries on the Volume 71 Index subfile. Most recall failures were due to the use of misspelled terms or faulty logic on the text search profile. In a few instances the search term derivation was incomplete.

Relevancy failures were classified according to the reason for failure. The three most significant types and their percentages are shown in Table IV. These reasons for failure are not mutually exclusive, however; there may be more than one reason for a given failure.

The most common type of relevancy failure were cases where the required structural unit was not present in the retrieved substance. In some of these cases the terms being searched for were embedded in larger, unwanted terms. For example, the unwanted "isothiazole" would be retrieved by *THIAZOL* along with the desired "thiazole". In searches of the CDB subfile these unwanted terms were sometimes negated using the CAS 360 Text Search System's cumulative weighting feature. The capability to safely negate unwanted terms was not available with the version of the experimental programs used to search the Volume 71 Index subfile. The smaller percentage of relevancy failures of the first type in the CDB searches is probably due primarily to this negation of unwanted terms. Cumulative weighting was used to "ignore" unwanted terms by assigning the desired term a positive weight and the unwanted term a negative weight of the same value. For example, in a profile with a threshold weight of +2 and the search terms

ISOTHIAZOL −2

OR *THIAZOL* +2

the Index name

$$\text{2-}\overbrace{\text{Thiazole}}^{+2}\text{carboxamide, 3-(3-}\underbrace{\text{iso}\overbrace{\text{thiazol}}^{+2}}_{-2}\text{yl)-}$$

total weight = 2

would be retrieved but

$$\text{3-}\underbrace{\text{Iso}\overbrace{\text{thiazol}}^{+2}}_{-2}\text{ecarboxamide}\qquad\text{total weight} = 0$$

would not be; the term "isothiazole" is in effect ignored.

Other instances of the first type of relevancy failure were cases where some of the structural units in the desired substructure were not encoded in order to simplify profile construction.

The second type of relevancy failure occurred where the query contained an unsubstituted ring or chain, but a substance which contained an unwanted substituent was retrieved. Since hydrogen is not explicitly expressed in most CA Index Names, it is usually not possible to search for it with a specific search term. Consequently, this type of irrelevant retrieval was difficult to prevent with the search system facilities used in this experiment.

The third type of relevancy failure resulted from false coordination of terms for two or more structural units in the desired substructure. AND logic is typically used between terms for two structural units because the terms are not necessarily contiguous in a CA Index Name due to ordering rules. This failure was not as prominent for the Volume 71 Index searches as it was for the CDB searches, probably because of the ability to specify, in the searches of Volume 71, the name segments in which the search terms can occur. This name segment or data element specification for search terms prevented some of the false coordinations with terms in different data elements.

Figure 14 relates the number of encoded nomenclaturally significant structural units for a query to the relevancy of the retrieval. Queries involving a single unit, primarily rings and functional groups, were the most precise, probably because a unit is usually defined by a single set of alternate terms; i.e., AND logic connections, which afford the possibility of false coordination, are not usually used. The mean relevancy was the lowest for the two-unit searches but increased for both the three- and four-unit searches. The higher relevancy for the three- and four-unit searches is probably due to increased restrictiveness in the search logic since each unit corresponds to an AND logic parameter.

As shown in Table III the relevancy of the nomenclature substructure searches was intermediate between the relevancy obtained for the fragment and iterative search (100%). There was also a significant difference between the amount of time needed for the profile development phase for the three methods of substructure searching. Development of a nomenclature profile typically required 1.5 hr including time consumed in using search aids. However, encoding times varied from about 5 min to the several hours required for one query that involved use of ring system support materials. The combined encoding time for a typical fragment and iterative search was approximately 0.75 hr; however, this time saving was usually offset by the additional time needed to "decode" the registry number output from the iterative search.

For the CDB searches, which involved the use of two formal, but different, search systems, the computer time (CPU) required for the nomenclature searches was a few seconds less than the time for the combined fragment and iterative searches on the same computer model.

SEARCH SYSTEM FEATURES

To improve the relevancy of nomenclature substructure searches, the batch search program used in initial experiments has been modified to include features which emulate some of the expanded logic facilities used by several other text search systems. An unwanted term which contains a desired term can now be "ignored" either directly or by using an added cumulative weighting feature. Another available feature allows a user to specify that two terms must occur in a specified order, i.e., that one term is followed by the other after an intervening character string of any length. Because of the syntax used in CA Index nomenclature, this capability is very useful in maximizing the relevancy of nomenclature substructure searches that are framed to achieve complete recall. This capability can be used to prevent irrelevant answers which contain the desired terms but not in the required sequence, i.e., which give false coordination of search terms. For example, the profile shown in Figure 12 could retrieve names for substances such as "Benzeneacetic acid, 4-(4-acetyl-2-thiazolyl)-3-(aminocarbonyl)-," and "2-Napthalenecarboxamide, *N*-(4-phenyl-2-thiazolyl)-" in which the ring and functional group are not attached. The first of these names could be prevented by specifying that the ring term follow the functional groups, e.g., *AMINO* followed by *CARBONYL* followed by *THIAZOL* in a search of the Substituent or Name Modification segments (see Figure 10a). The second name could be prevented by specifying that the functional group term follow the ring term, e.g., *THIAZOL* followed by *CARBOXAMIDE* (see Figure 10b and 10c).

The capability to specify the maximum number of intervening characters between an ordered pair of terms was also included in the program upgrading. This feature allows an even more precise profile than the "followed-by" logic feature. For example, the "followed-by" specification would not prevent the retrieval of an irrelevant name such as "Benzoic acid, 4-[4-amino-5-(bromocarbonyl)-2-thiazolyl]-", for the sample query. This "false drop" could be prevented by specifying that no more than one character occur between the *AMINO* and *CARBONYL* terms. Allowing only one character permits an enclosing mark after "amino".

Our current experimental batch search programs also include a procedure that examines the enclosing mark pattern between an ordered pair of terms to determine if the structural units described by the terms are connected to each other. When two terms in a CA Index Name describe connecting structural units, there is a definite enclosing mark pattern between these terms in the name segment string. For example, in a Substituent segment of a substitutive name, two nomenclature terms can usually be assumed to describe connecting units if the second term is directly followed by a right enclosing mark and if between the terms there are matched pairs of enclosing marks, a single right enclosing mark directly following the first term, or no enclosing marks. Figure 15 shows some examples of substitutive names with terms describing connecting and nonconnecting units.

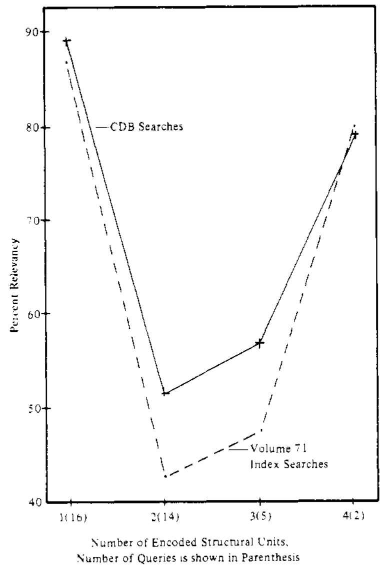


Figure 14. Relevancy of retrievals from experimental queries.

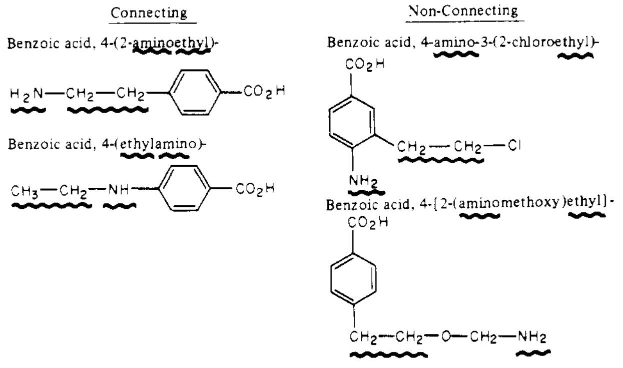

Connecting
Benzoic acid, 4-(2-aminoethyl)-
H2N—CH2—CH2—
—CO2H
Benzoic acid, 4-(ethylamino)-
CH3—CH2—NH—
—CO2H
Non-Connecting
Benzoic acid, 4-amino-3-(2-chloroethyl)-
CO2H
CH2—CH2—Cl
NH2
Benzoic acid, 4-[2-(aminomethoxy)ethyl]-
CO2H
CH2—CH2—O—CH2—NH2

Figure 15. Substitutive names with connecting and nonconnecting units.

To date most of our nomenclature substructure searches, both 8CI and 9CI, have been batch mode, serial searches performed at CAS. We have, however, performed several searches on the CHEMLINE Retrieval File Set[8,9] using their on-line, interactive search system. Each substance represented in a unit record of this file set has one or more CA Index Names and usually several synonymous names such as acronyms, trade names, natural product names, and other systematic names which are not used as CA Index Names. Wiswesser Line Notations (WLN) are also included for many of the substances. Two types of search are possible with this file: (1) an index or inverted file search in which a file of index terms derived from the substance names, molecular formula, CAS Registry Numbers, or WLN's can be searched to identify the appropriate unit records and (2) a string search in which the substance names, molecular formulas, CAS Registry Numbers, or WLN's in the unit records retrieved in the index search can be further searched to improve the relevancy of the retrieval. Left truncation is available in the string search but not in the dictionary search. The system also has a capability which allows the user to "browse" through the terms in the inverted file. This is useful in determining the proper truncation.

In addition to the well-known advantages of interactive over batch processing, the use of an interactive system such as the one used in CHEMLINE searching often simplifies nomenclature substructure searching. With CHEMLINE it is usually possible to vary the sequence in which the terms are used and the stage in the search at which the logic is applied. Since each structural unit can be searched for separately, the entire search can be terminated if, after a search for a given unit, no names are retrieved or the set of names retrieved is small enough to review.

Another useful feature of the CHEMLINE file set organization is that each element symbol is indexed individually, thus allowing for a molecular formula screen to be applied prior to the nomenclature search. This molecular formula search capability is particularly useful when the desired substructure contains several heteroatoms or a relatively rare heteroatom. For example, suppose a desired substructure contains a selenium atom. Since this atom is relatively rare, an initial search for SE may result in no retrieval or a small amount of retrieval, thus eliminating the need to perform the nomenclature search.

COMPARISON OF 8CI AND 9CI SEARCHING

The simplification and standardization of CA Index Name selection practices that were made for the 9CI period are advantageous for nomenclature substructure searching in several respects. The elimination of many nonsystematic names in favor of the corresponding, fully systematic names results in a reduction of the number of search terms needed in a profile. We have found that about 15% of the terms of the 37 8CI profiles would not be needed for 9CI nomenclature searching. This reduction is also illustrated by the fact that while the 8CI Nonsystematic Name Guide contained about 2500 terms this number was reduced to about 1400 in the 9CI version. The smaller number of search terms causes a corresponding reduction in the time needed to encode a profile and in the computer time needed to run the search. Typically, our 9CI profiles have contained between 10 and 30 unique terms.

To determine if there is a significant difference in the search performance between searching files of 8CI names and searching files of 9CI names, we have performed several substructure searches on corresponding files of 7980 8CI and 9CI names. The 8CI and 9CI support materials were used in developing the profiles. We found no significant difference between the files in either recall or relevancy.

CORRELATIVE SEARCHES WITH NONSUBSTANCE DATA

The relevancy of a nomenclature substructure search can sometimes be improved, or profile development simplified, by coordinating the nomenclature search with a search for nonsubstance, or concept terms. Searches of this type are useful for identification and prediction of relationships between classes of chemical substances and their properties or biological activities.

Improved relevancy is likely because irrelevant substances will not be retrieved unless the nonsubstance data associated with them also satisfies the search. Thus, it is usually possible to achieve both high recall and high relevancy for the substructure aspect when performing a correlative substructure/concept search.

In the substructure/concept searches performed at CAS, we have usually framed the nomenclature portion very generically, as in the profile for the sample query (Figure 12). This strategy has resulted in simplified nomenclature profile construction while still leading to good relevancy for the substructure aspect because of the restrictiveness of the nonsubstance portion.

Table V. Using Search Screens for Sample Query

Screen type	% file screenout
Molecular formula character	88
Nomenclature number-of-characters	24
Nomenclature character	76
Nomenclature search term	96
Combination of all four screens	98.6

SCREENING TECHNIQUES

To minimize the amount of computer time needed in the batch mode, serial searching for substructures via nomenclature at CAS, we have employed several screening techniques. These screens reduce the number of records for which a complete nomenclature search is needed, i.e., the number of records which are searched with all the nomenclature terms. These screening techniques have resulted in considerable savings in computer search time, especially in the searching of large retrospective files such as several volumes of the CA Integrated Subject File.

Table V shows the percentage of file screenout obtained when several types of screens were used in a search of an issue of CA Subject Index Alert (CASIA) (41,538 nomenclature records) for the sample query (Figure 2).

Molecular Formula Screens. A search of the molecular formula data for the presence of characters in the required element symbols can sometimes considerably reduce the file on which a nomenclature search is performed; i.e., only the names of substances whose molecular formulas contain the required characters need be searched. This is especially true if the desired substructure contains a relatively infrequently occurring heteroatom. For example, only about 20% of the substances on the CAS Chemical Registry files have a sulfur atom. Thus, searching the molecular formula and determining the presence or absence of an "S" should make a very effective screen. In the sample search for "thiazolecarboxamide", a molecular formula scan for "C", "H", "N", "O", and "S" resulted in 88% screenout; i.e., only 12% of the substances had a molecular formula with these required element symbols.

In our current search programs, searching for a character or group of characters in a search field is accomplished using the PL/1 programming language function, VERIFY. This function "verifies" the presence of a user-supplied character or set of characters in the data.

Another procedure available in our programs determines if the molecular formula data contains a "minimum molecular formula" for the desired substructure, i.e., the symbols and their minimum "counts" that must be present in a molecular formula of any relevant substance. Since it requires more computer processing time than molecular formula verification, the minimum molecular formula routine, when used, is called after the verification step to further reduce the file.

Nomenclature Number-of-Characters Screens. The number of characters in a CA Index Name or a specific nomenclature data element can also be used as a nomenclature search screen. This can be accomplished by comparing the number of characters in the data to a user-specified minimum number of characters which every relevant name must contain. The nomenclature search is then performed only on names (or name segments) containing the minimum number of characters. For example, in the sample search for "thiazolecarboxamide" (Figure 2), a substance name with only a Heading Parent segment would need to contain at least 12 characters to be relevant, e.g., "Bacitracin F". If the name also consisted of a Substituent or Name

Modification segment, it would need to contain at least 29 characters, e.g., "2-Thiazolecarboxamide, *N*-ethyl-". In the sample search, the number-of-characters screen resulted in 24% screenout. (See Table V. The average number of characters in a CA Index Name in the issue of CASIA used was found to be 54.3.)

Number-of-characters screens for systematic names are usually more effective when nomenclature terms for several structural units are used, since relevant answers would contain a name term(s) for each structural unit and thus the minimum number of characters would be relatively large.

Nomenclature Character Screens. As with molecular formulas, nomenclature data can also be scanned for set(s) of characters which a potentially relevant name must contain. The characters in these sets are common characters in the nomenclature search terms. If these characters are found in a name, a complete nomenclature search is performed on that name. For example, with sample question (Figure 2) every relevant answer should contain one of the following alphabetic character sets: "A, B, C, H, I, M, N, O, R, T"; "A, H, I, O, T, X, Z"; or "A, B, C, I, N, R, T". The first two sets were derived from the systematic name terms whereas the last set contains the unique characters in the nonsystematic term "Bacitracin". In the sample search, these character screens resulted in 76% screenout (see Table V).

Nomenclature Search Term Screens. Another screen that can be used in nomenclature substructure searching is to search initially for only one of the structural units in the desired substructure, usually the one with the smallest number of search terms; the names retrieved by this partial search are then searched to locate substances containing the remaining structural units. In the sample search, the "carboxamide" unit was used as a screen with a screenout of 96% (see Table V).

In correlative nomenclature/nonsubstance searching a required search term set for either aspect could be used.

Screen Combinations. Although there is considerable overlap among the screens used, applying them in combination produces an overall screen that is more effective than the individual screens. For the sample search, the combination of the molecular formula and nomenclature character screens resulted in a 94.6% screenout; the combination of the molecular formula, nomenclature character, and nomenclature number-of-characters resulted in a 94.6% screenout; and the combination of molecular formula, nomenclature character, nomenclature number-of-characters, and "carboxamide" search terms resulted in a 98.6% screenout.

CONCLUSIONS

The recall and relevancy of the retrieval in nomenclature substructure searching depends on several factors: the search system characteristics, the nature of the query, the nomenclature background of the encoder and his encoding strategy and technique, the amount of time available for profile construction, and the availability of support materials.

The results obtained in the substructure searches performed at CAS indicate that profiles for nomenclature substructure searches can be framed on a complete recall basis by an encoder with a limited background in CA Index nomenclature, provided that support materials such as those described in this paper are available and that the search system has at least minimum capabilities, i.e., those assumed in encoding the sample query. Even without specialized search aids, nomenclature profiles that will obtain very high, but not necessarily complete, recall can be developed for most substructure queries. However, for a few types of queries it may not be practical or may be too time-consuming to frame a profile on a complete recall basis. One example is a ring query in which an envelope ring (i.e., a ring which contains bonds that divide it into two or more rings) is considered an acceptable answer, for example, "quinoline" as a desired retrieval for the query, "all ten-membered rings." The KWOC Index does not provide straightforward access to "quinoline" as a ten-membered ring.

Another type of query is a highly generic substructure question with a small number of atoms such as "C–C or C–N, cyclic or acyclic." These substructures are "smaller" than most substructures that are named in CA Index Names and, consequently, require a significant number of nomenclature terms to identify them. It is much easier to search for such substructures only via the molecular formula, although the retrieval may be somewhat less precise.

The relevancy that can be obtained in nomenclature searches depends on factors cited above but also to a large extent on whether complete recall is a requirement. For most queries, profiles that will result in high relevancy and high, but not necessarily complete, recall are relatively easy to construct if the search system has the minimum capabilities. When complete recall is desired, the nature of the query becomes important. Our experiments indicate that searches for a substructure corresponding to a single, nomenclaturally significant unit such as a ring or functional group will usually result in high precision, while the retrieval for a two- or three-unit query will usually be less precise.

To obtain very high relevancy in nomenclature substructure searching without adversely affecting recall usually requires a good knowledge of CA Index nomenclature rules and/or a search system with expanded logic facilities. However, if the search is a correlative substructure/concept search, high relevancy can be obtained by virtue of the restrictiveness of the concept portion.

Our experiments indicate that the time required to construct a nomenclature substructure search profile will vary considerably, with the most time-consuming step being the derivation of the corresponding nomenclature terms. In general, functional group and isolated ring system terms are the easiest to derive. The derivation of terms for embedded ring systems and terms for generic substructures which do not correspond directly to a nomenclaturally significant unit are the most difficult.

In conclusion, the use of CA Index nomenclature and its supporting data as a basis for substructure search with computer programs appears to have both advantages and disadvantages. Some possible disadvantages are:

1. Complete relevancy is difficult to obtain for queries framed on a complete recall basis.
2. The time required to encode a profile is very sensitive to the nature of the query and can vary considerably from one query to another.
3. Substructures that do not consist of nomenclaturally significant structural unit(s) are sometimes difficult to encode; i. e., the profile development can require considerable time.
4. Support materials are usually required to frame a query on a complete recall basis.

Some possible advantages are:

1. Certain classes of substances which cannot easily be defined in terms of a precise substructure can readily be retrieved in nomenclature searching. For example, steroid systems including the various cyclo, homo, nor, and seco systems can be retrieved by using a small number of steroid stem terms.
2. Nomenclature substructure searching can be implemented on a "ready-made" system, i.e., text search systems. Batch mode, text search systems are available at many of the information centers that process computer-readable chemical information, and, recently, services using on-line, interactive systems with remote text terminal access have been made available.

3. Several computer-readable data bases and files containing CA Index Name and related data are currently available. Currently available from CAS are: Chemical-Biological Activities, Polymer Science & Technology, CA Integrated Subject File, CA Subject Index Alert, Energy, Materials, Food and Agricultural Chemistry, and Ecology and Environment. Also available is CHEMLINE which is part of the TOXLINE service provided by the National Library of Medicine.

Most of the above files and data bases also contain nonsubstance data, and this makes substructure/concept correlative searches possible.

4. Search aids which assist the encoder in developing effective nomenclature substructure searches are available from CAS.

5. The simplification and standardization of nomenclature practices for the 9CI period have greatly simplified profile construction and reduced the computer time (batch mode) needed for nomenclature substructure searching.

6. Index nomenclature and the supporting molecular formulas contain "built-in" screens which can be used to reduce the number of records in batch mode, serial searching which need a "full" nomenclature search.

7. The CA Index Name retrieval in nomenclature substructure search can readily be decoded to determine the relevance of the corresponding substances. This is possible because almost all CA Index Names are systematic and because the family of relevant names has already been anticipated in the encoding process.

8. In correlative substructure/concept searches the lack of precision in the nomenclature searches is usually offset by the restrictiveness of the concept portion.

ACKNOWLEDGMENT

The development of the techniques and support materials for 8CI nomenclature substructure searching was partially supported by the National Science Foundation and completed in 1971. CAS gratefully acknowledges this support.

LITERATURE CITED

(1) "Chemical Abstracts Service Search Aids for the 9th Collective Index Period (1972–1976)," Chemical Abstracts Service, Columbus, Ohio, June 1974, 109 pp, ISBN 8412-0198-6, LCN 74-80986.

(2) Chemical Abstracts Service Search Aids for Substructure Searching of 8CI Computer-Readable Chemical Nomenclature Files, Chemical Abstracts Service, Columbus, Ohio, May 1973 (available on microform through the National Technical Information Service, Springfield, Va., PB 229578).

(3) Donaldson, N., Powell, W. H., Rowlett, R. J., Jr., White, R. W., and Yorka, K. V., "*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.

(4) "Substructure Search (Background Information and Question Coding Instructions)," Chemical Abstracts Service, Columbus, Ohio, Aug 1970, 115 pp.

(5) "Substructure Searching of Computer-Readable CAS 9CI Chemical Nomenclature Files (Based on Nomenclature in the Ninth Collective Index of Chemical Abstracts) (1972–1976)," Chemical Abstracts Service, Columbus, Ohio, Aug 1974, 128 pp, ISBN 8412-0204-4, LCN 74-14778.

(6) "Substructure Searching of Computer-Readable CAS 8CI Chemical Nomenclature Files (Based on Nomenclature in the Eighth Collective Index of Chemical Abstracts) (1967–1972)," Chemical Abstracts Service, Columbus, Ohio; May 1973, 162 pp, ISBN 8412-0182-X, LCN 73-84537 (available through the National Technical Information Service, Springfield, Va., PB 229578).

(7) "Text Searching (Documentation for Searching the CAS Files Using the IBM S/360 Operating Systems)," 2nd ed., Chemical Abstracts Service, Columbus, Ohio, 1969, 41 pp.

(8) Vasta, B. M., "TOXLINE Chemical Dictionary File," presented to the 167th National Meeting of the American Chemical Society, Los Angeles, Calif., April 2, 1974.

(9) Vasta, B. M., Walker, D. F., Jr., Schultheisz, R., and Hummel, D. J., "Conversion of TOXLINE Chemical Dictionary into CHEMLINE," presented to the 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 11, 1974.

(10) Wigington, R. L., "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.