Menis, Dr. O.
Analytical Chemistry Division
National Bureau of Standards
U.S. Department of Commerce
Washington, D.C. 20234

Neurath, Dr. Hans
Department of Biochemistry
School of Medicine
University of Washington
Seattle, Washington 98105

Noland, Dr. Wayland E.
Department of Chemistry
University of Minnesota
Minneapolis, Minnesota 55455

Paul, Dr. Martin A.
1772 Horatio Avenue
Merrick, New York 11566

Reynolds, Miss Helen L.
The Association of Official Analytical Chemists
Box 540
Benjamin Franklin Station
Washington, D.C. 20044

Skolnik, Dr. Herman
Hercules Inc., Research Center
Wilmington, Delaware 19899

Smith, Dr. Peter A. S.
Department of Chemistry
University of Michigan
Ann Arbor, Michigan 48104

Spindel, Dr. William
Division of Chemistry and Chemical Technology
NAS-NRC 2101 Constitution Avenue
Washington, D.C. 20418

Stocker, Dr. Jack H.
Department of Chemistry
University of New Orleans
Lakefront
New Orleans, Louisiana 70122

Walling, Dr. Cheves
Department of Chemistry
University of Utah
Salt Lake City, Utah 84112

White, Dr. Robert W.
Chemical Abstracts Service
P.O. Box 3012
Columbus, Ohio 43210

Young, Dr. John A.
U.A.G.
Paseo de las Aguilas 7000
Lomas del Valle
Guadalajara, Jalisco
Mexico

Gutsche, Dr. C. David, Liaison Member
Department of Chemistry
Washington University
St. Louis, Missouri 63130

Structure-Text and Nomenclature-Text Searching for Chemical Information: an Experiment with the *Chemical Abstracts Integrated Subject File* and Registry System

J. F. B. ROWLAND[*] and MARGARET A. VEAL

United Kingdom Chemical Information Service, University of Nottingham, Nottingham NG7 2RD,
United Kingdom

Received February 7, 1977

The *Chemical Abstracts Integrated Subject File* (the computer-readable file corresponding to the volume subject index of *Chemical Abstracts*) and a subset of the *Chemical Abstracts* Registry System file of connection tables have been used for a comparative test of nomenclature-text and structure-text searching. The structure search used a bit-mask searching technique, with a fragment code derived from the connection tables by the statistical methods of M. F. Lynch's group at the University of Sheffield. The same test questions were used to search corresponding nomenclature-text and structure-text data bases. The conclusions were that the fragment bit-screen search was capable of providing a search service of adequate quality even without an atom-by-atom search, that the fragment search performed on balance with superior cost-effectiveness to the nomenclature search, that the nomenclature search nevertheless gave a quality of performance (recall and precision) that would be regarded as acceptable in a commercial information-retrieval service, and that, within the fragment code, bond-centered fragments gave a superior performance to atom-centered fragments.

INTRODUCTION

The volume subject indexes to *Chemical Abstracts* refer to compounds by means of highly systematic nomenclature, the version used from 1972 to 1976 being Ninth Collective Index (9CI) nomenclature.[1] The computer-readable data base corresponding to these volume subject indexes is known as the *Chemical Abstracts Integrated Subject File* (CAISF).[2] Like the volume subject indexes themselves, CAISF is organized in two parts, the General Subject file and the Chemical Substance file; within each file, the arrangement is alphabetical upon the parent headings. The file therefore differs from most computer-readable information files in that the different pieces of information referring to a document are not found together, but are scattered throughout an alphabetical index.

This file enables information scientists to carry out substructure searches, using the systematic nomenclature as the structure file, and link them to searches of general concepts

[*] To whom correspondence should be addressed.

Table I

Chemical substance entries	General subject entries
Heading Parent	Concept Heading
Homograph Definition	Homograph Definition
Line Formula	Qualifier
Substituent	Functional Category
Name Modification	Text Modification
Stereochemistry	CA Publication Citation
Qualifier	
Registry Number	
Functional Category	
Text Modification	
Preferred Order Molform	
CA Publication Citation	

in the remainder of the file. Furthermore, the file contains *Chemical Abstracts* Registry Numbers for the compounds indexed. Thus one can obtain from the CAS Registry System[3] the connection tables of the compounds mentioned in one volume of CA. From such a subset of the Registry System, one can derive a file of bit-masks based on structural fragments. The scope therefore existed for a comparative study of the effectiveness of a structure-text search with the CAISF nomenclature files as the structure data base, and a parallel structure-text search with a fragment bit-mask file as the structure data base.

Lynch and his co-workers[4-8] have developed a sophisticated technique for deriving a fragment code from a connection-table file, by which one selects the fragments for inclusion in the code on the basis of their observed frequency of occurrence in a representative sample of the connection-table file itself. Lynch's methods have not previously been tested in an operational or near-operational environment. Furthermore, Lynch believes that fragments centered on a bond, rather than the more conventional atom-centered fragments, will give superior retrieval performance.

We have carried out a comparative test of nomenclature-text and structure-text searching. The same data base—CAISF Volume 76 and the compounds mentioned in it—was used for both parts of the test, and in the structure searching we used screens derived by Lynch's methods. We tested atom-centered and bond-centered fragments comparatively, to test Lynch's hypothesis that the bond-centered ones would give superior performance. The same file of test questions, obtained from research chemists working in industry and in universities, was used for all the searches.

A fuller description of this work has been written in the form of a report to the British Library Research and Development Department[9] and can be obtained on microfiche from the British Library Lending Division, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, U.K.

METHODOLOGY

The CAISF appears once every 6 months and is divided into two parts, the General Subject and Chemical Substance files;[2] the data elements that may be present in any index entry are listed in Table I. Each document indexed gives rise to at least one index entry; on average each document has just over four Chemical Substance entries and just over two General Subject entries.

Table II. Examples of Index Entries

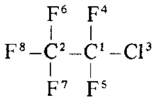
Data-element type	Entry
(a) Chemical Substance Entry	
Registry number	023150303
CA abstract citation	CA07111050487A[a]
CA section number	CA03400
Heading parent	Serine
Substituent	3-[*p*-(Methylthio)phenyl]-
Stereochemistry	D-*threo*-
Molecular formula	Cl0H13NO3S
(b) General Subject Entry	
CA abstract citation	CA07112052907M[a]
CA section number	CA05600
Concept heading	Springs
Homograph definition	Mechanical
Text modification	Bronze, mech. properties of thermomech. textured phosphor

[a] The two citations mean CA Vol. 71, issue 11, abstract 50487 (check digit A) and CA Vol. 71, issue 12, abstract 52907 (check digit M).

Table III. A CAS Registry II Connection Table

Reg. No 76-15-3
TEXT NS

Atom no.	1	2	3	4	5	6	7	8
Conn.		1	1	1	1	2	2	2
Element	C	C	CL	F	F	F	F	F
Bond		−1	−1	−1	−1	−1	−1	−1
H-count								



For this experiment we used CAISF Volume 76, the index file corresponding to Volume 76 of *Chemical Abstracts*, January to June 1972. This was the first volume to be included in the Ninth Collective Index, which covers 1972 to 1976 inclusive. Also, *Chemical Abstracts Service* prepared for us a file of connection tables of all the compounds indexed in Volume 76; this was possible because the CAISF Chemical Substance file contains Registry Numbers. The Registry subset file thus created was used in the bit-mask search phase of the project; it contained the connection tables in Registry II format. Tables II and III show examples of entries in our two test files, CAISF and the Registry subset, respectively.

Using these data bases, we devised the comparative experiment as follows. Programs were written to take the incoming CAISF and to divide this into three files: a nomenclature file, in which each entry was the systematic name of a specific compound identified by its Registry Number; a text file, containing the remainder of the information from CAISF, each entry being identified by its CA Abstract Number; and a Registry Number–Abstract Number concordance. The two main files (the nomenclature and text files) were inverted to give search and dictionary files under the United Kingdom Chemical Information Service (UKCIS) inverted-file retrieval system, INFIRS.[10] Figure 1 is a summary flowchart of the nomenclature-text system.

In order to restrict the search and dictionary files to a manageable size for UKCIS's computer (an ICL System 4/50, approximately equivalent to an IBM 360/40, with 128K bytes of core store), it was necessary to make some modifications to the standard INFIRS programs. In INFIRS, each word that is listed (the "key") is followed by a list of the addresses in the file of documents where the word is found. The address is normally in the form "document, sentence, word"; in a *Chemical Abstracts Condensates* file, for example, a "sentence" is the title or one of the keyword phrases. In the nomenclature file, we decided to use addresses of the form "registry number, word"; so, within each compound, each data element (parent heading, substituent, etc.) was regarded as one word, and was numbered successively through the compound, with the exception of the name modification, where the separate words were numbered separately. This arrangement enabled us to search for the different parts of a systematic name in order, an important requirement in nomenclature searching. The distinction between the different data elements (parent heading, substituent, etc.) was maintained to aid in profile construction. We also decided to install a permuted dictionary for the nomenclature file; this option

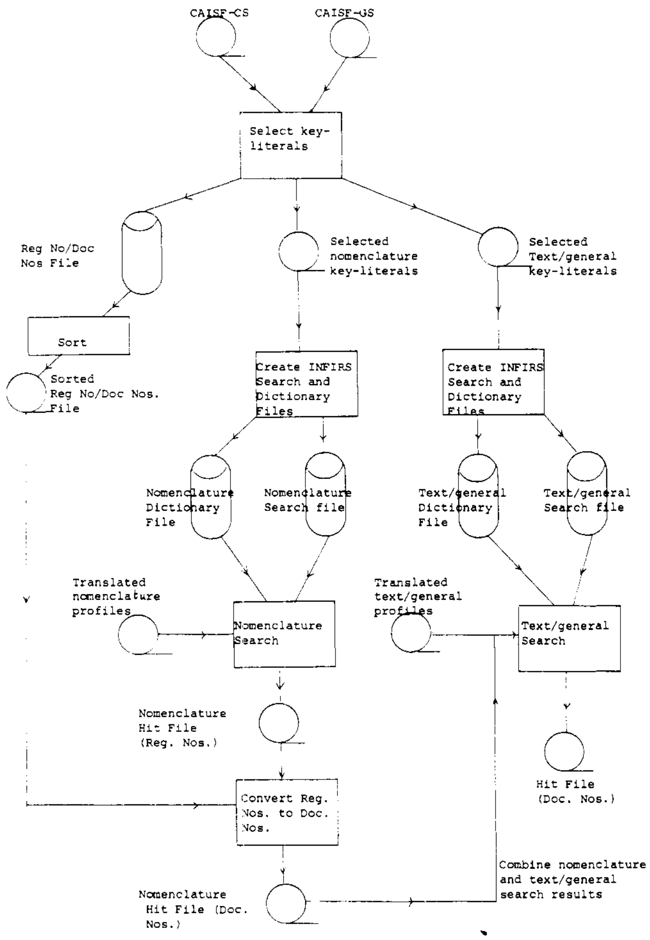

CAISF-CS
CAISF-GS
Select key-literals
Reg No/Doc Nos File
Selected nomenclature key-literals
Selected Text/general key-literals
Sort
Sorted Reg No/Doc Nos. File
Create INFIRS Search and Dictionary Files
Create INFIRS Search and Dictionary Files
Nomenclature Dictionary File
Nomenclature Search file
Text/general Dictionary File
Text/general Search file
Translated nomenclature profiles
Nomenclature Search
Translated text/general profiles
Text/general Search
Nomenclature Hit File (Reg. Nos.)
Hit File (Doc. Nos.)
Convert Reg. Nos. to Doc. Nos.
Combine nomenclature and text/general search results
Nomenclature Hit File (Doc. Nos.)

Figure 1. Outline system flowchart; nomenclature and text search.

in INFIRS means that the dictionary file entries for (e.g.) SULFURIC would be SULFURIC, ULFURIC, LFURIC, and so on down to IC, enabling the searcher to employ leading truncation. The dictionary is made very bulky by this, but in nomenclature searching it is essential to be able to find embedded terms.

By contrast, we decided to install a much simplified version of INFIRS to search the text file. Here we used only a "document" address in the inverted file, with no indication of each word's position within the document. There were two reasons for this: (a) by this means the file was shortened, as the addresses were shorter, and duplicate occurrences of the same word within the same document were automatically eliminated; (b) in an index file it would be difficult to assign "sentence" numbers, as the different "sentences" (index entries) referring to a particular document are not together on the file. In the text file, we did not permute the dictionary, and thus no leading truncation was possible. This again reduced the size of the file very substantially. Finally, in the text file, we did not maintain any distinction between the different data elements; thus in search profiles the term type "text" was used for all terms in the text file. So, in this file, all that was available for search was Boolean "and", "or", and "not" logic, at the document level, and trailing truncation. This simple search was, however, sufficient, since in this project we were essentially aiming to answer structure-based questions and using the text as a tool for sharpening precision; the text search was not intended to stand alone.

The test questions that we used for this experiment were selected from two collections. One was a collection of questions used in a project at the Experimental Information Unit, University of Oxford, with which one of us (J.F.B.R.) was previously associated;[11,12] these were from university and industrial chemists, with university workers predominating. In the Oxford experiment, a live S.D.I. service based on the *Index Chemicus Registry System* data base had been provided to the users, and there were therefore records of the actual performance of these users' questions. The second collection was one solicited by UKCIS for its structure-search research, which mostly came from industrial chemists. This collection has not been used for any real-life service to users. The questions used in the experiment were selected from these two collections; the choice was made on the basis that we required structure-text questions, and so we chose those with a significant conceptual as well as the substructure requirement. Table IV lists brief statements of the questions used in the experiment.

NOMENCLATURE SEARCH

The nomenclature search was performed essentially according to principles established by Fisanick et al. at CAS.[13] The search aids published by CAS were also consulted;[14] it was found that the principles for composing a satisfactory nomenclature search profile were also similar to those that one would use in a Wiswesser Line Notation search.[11]

In CAS systematic nomenclature,[1] functional groups are arranged in an order of precedence, and the senior group present is named as a suffix; the skeletal group to which this senior functional group is attached as suffix becomes the heading parent. Examples of heading parents are "benzenamine", "2-butenoic acid", "1,4-dioxan-2-one". Most of the remaining parts of the molecule are then given in the "substituent" data element, which might for example contain entries such as "3-chloro-5-methyl". Certain classes of compounds, most notably esters, are named as derivatives; thus an ester of a phosphorothioic acid would have the ester groups in the "name modification" field rather than the "substituent" field.

These features of CAS nomenclature make for some difficulty in searching for substructures. One has to allow for the required structure being in the parent (with the senior function named as a suffix), in the substituent, in the name modification, or, for large substructures, partly in one data element and partly in another. All these possibilities have to be allowed for in the profile.

Furthermore, one also has to allow for the required substructure being in a compound that is still named nonsystematically. Although the current version of CAS nomenclature, Ninth Collective Index nomenclature, has far fewer nonsystematic names than earlier versions, a number of important classes of compounds are still not systematically named. The difficulty is not so much where the search is actually for one of these groups, e.g., a search for 17-hydroxy steroids (in this case one allows for the semisystematic nature of the nomenclature), it is where one requires a substructure (e.g., hydroxycyclopentanes) that might be buried in a structure named systematically, but might also be buried in one named non- or semi-systematically. Thus the search profile has to contain both systematic and nonsystematic terms and, in compiling the profile, one has to predict all the nonsystematically named types of compounds in which the fragment might appear. This is difficult, and a large proportion of the work in compiling the profile is thus necessitated; the proportion of documents retrieved by the nonsystematic terms, however, may be low. It would be in the interests of those who wish to perform substructure searches via nomenclature if CAS were to make their nomenclature system even more systematic than it is today; such a move, however, would probably be unpopular with users of the printed CAS indexes.[15,16]

Using the search terms selected as outlined above, we then wrote search profiles of the normal INFIRS type[10] to search

Table IV. Questions Used in This Project

Question identifier	Brief statement of the query
	Oxford Questions
SOAI1	Pyrylium salts and pyrandiones and their use in the synthesis of other heterocycles
SOCI1	Cannabis constituents
SODB1	Reserpine, isoreserpiline, ellipticine, santonin, and other alkaloids from *Ochrosia, Excavatia,* and *Apocyanaceae*
SODE1	X-ray, UV, MS, NMR, and IR studies of local anaesthetics
SODG1	Plant steroids and sapogenins as a source of material for steroid drugs and contraceptives
SOFC1	Total synthesis of chorismic acid and *cis*-decalin sesquiterpenes; mechanism of biosynthesis of phenylalanine from chorismic acid; vitamin B_{12}
SOFQ1	Mass spectrometry of ergot alkaloids, sesqui- and diterpenes, hexahydrophenanthrenes and ^{13}C-labeled compounds; GC–MS; photoionization; substituent effects; metastable spectra
SOGD1	Thiazolopyrimidines and other chemical antineoplastic agents
SOGE1	Natural and model membranes; interaction between cholesterol and phospholipids, effects of analgesics and antibiotics on membranes; study of membranes by IR (ATR), polarography, and cyclovoltammetry
SOGG1	Fungal and microbiological hydroxylation of steroids and other alicyclics; autoxidation of steroids, steroid hydroperoxides
SOHA1	Nucleic acids (fractionation and structure assignment); purines and pyrimidines; interferon
SOHD1	Stereochemistry of sulfoxides; cyclic acetals of carbohydrates
SOHL1	Pyrrole-2-carboxylates; pyrrolo[2,1-*c*]-1,4-oxazine-1-ones; pyrrolo-[1,2-*a*]pyrazines; indolyl[2,1-*c*]-1,4-oxazin-1-ones; azetidin-2-ones
SOHP1	Crystallization of polymers; properties of PVC; ABS terpolymers
SOHQ1	Five-membered heterocycles containing two or three O, S, or N atoms, and carbonyl groups on the ring
SOHT1	Organic phosphorus oxyesters, especially cyclic ones, useful as antioxidants in polymers; organic hydroperoxides
SOKD1	NMR of fluorine compounds; three-membered ring compounds
SOSA2	Biological properties of selenophenes, bithienyls, fungicidal isocoumarins; fungal and plant benzophenones and anthraquinones; plant cardiac glycosides and their physiological effects
SOSB2	NMR, IR, and ORD studies of the conformations of amino acids and peptides; NMR studies of furans, thiophenes, pyrroles, piperidines, pyrimidines, pyrazines, piperazines, and pyridazines
SOVC2	*Rutaceae* alkaloids and other quinoline alkaloids
SOVH2	Pyridines, thiazoles, and isothiazoles substituted with NO_2 and another N substituent; thiadiazoles with either NO_2 or other N substituent; antimicrobial effects of these
SOVN2	Biosynthesis of thiamine (vitamin B_1); also several related pyrimidines, thiazoles, and 1,2,4-triazines
SOWC2	Coumarins, chromones, and other compounds occurring in *Guttiferae*
SOWE2	Quinolizines with CNS activity; synthetic routes to apomorphine; various fused-ring quinolizines
SOXB2	Redox potentials and polarography of triphenylmethyl dyes
SOZA2	Long-chain aliphatic compounds from tubercule lipids
	UKCIS Questions
SUCB2	Sulfobetaines useful as detergents
SUCB3	Fluorescence spectra of 1,3-diaryl-Δ^2-pyrazolines
SUCB4	Aliphatic hydroperoxides and their decomposition to alcohols and ketones
SUDA1	Biological, especially insecticidal, properties of cyclopropanecarboxylic acid esters
SUGA3	*p*-Nitrobenzyl esters of carboxylic acids, useful as a protecting group for the carboxylic acid group
SUGA4	2,4- or 2,6-Dinitrophenols substituted at the 4 or 6 position with a carbon substituent; pesticidal uses of these
SUGB1	UV spectra of cyclic β-unsaturated sulfides, sulfoxides, and sulfones
SUHA1	Mass spectrometry of methyl octadecadienoates
SUHA2	Mass spectrometry of Δ^4-steroids with one or more hydroxyl groups, one or no keto and one or no carboxyl group
SUHA4	Mass spectrometry of 9,10-anthraquinones with some hydroxyl or methoxyl groups
SUHA5	Mass spectrometry of partially methylated hexopyranosides
SUKA1	Biological activity of substituted quinazolones
SUKA2	Biological activity of carbazole derivatives saturated at 4a–9a
SULA1	Biological activity of morpholinomalonic acids and their esters
SULA2	Biological, especially CNS, activity of azepinoindoles
SULA3	Biological activity and synthesis of phenethylbiguanides
SULA4	Antibacterial activity of *cis*-glycidic acid and its esters

for the required compounds. A sample profile is shown in Table V. A number of subquestions, each looking for a different substructure, can be included in each profile; these are terminated by an "end subquestion" line and are lettered sequentially through the alphabet. Other lines in the profiles are search terms, identified as parent heading, substituent, etc. (see Table I), and logical lines to define required combinations of search terms. The output from the nomenclature search was in the form of Registry Numbers.

Table V. Question Amend - Amendment Report

	Profile no. SUCB3NM1	
Line no.	Search term type or operation	Operand
1	PARENT HEADING	PYRAZOL*
2	SUBSTITUENT	*PYRAZOL*
3	SUBSTITUENT	4,5-DIHYDRO*
4	SUBSTITUENT	1,3-DIPHENYL
5	PARENT HEADING	BENZ*
6	SUBSTITUENT	*PHEN*
7	LOGICAL	1,2 AND 3 S
8	LOGICAL	4 AND 7 S
9	LOGICAL	2 AND 3 S
10	LOGICAL	5 AND 9 S
11	LOGICAL	6 AND 10 S
12	LOGICAL	8, 11 OR
13	ENDSUBQUEST	A 200
END SUCB3NM1		

For each question, a text profile was also written to search for the conceptual part of the requirement. In the text profile, dummy search terms were included to represent the output of the nomenclature search. These dummy search terms appeared first in the profile, with term type "document number"; the search term is the subquestion letter from the appropriate subquestion in the nomenclature profile. Hits from the nomenclature search were first passed through a program which used the Registry Number/Abstract Number concordance file (produced earlier) to replace the Registry Numbers by the Abstract Numbers of the documents in which the compounds appeared. They were then input to the text

Table VI. Question Amend - Amendment Report

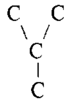
Line no.	Profile no. SUCB3TX1 Search term type or operation	Operand
1	DOCUMENT NUMBER	A
2	TEXT	FLUORESCEN*
3	LOGICAL	1 AND 2
4	ENDSUBQUEST	A 200
5	LOGICAL	1 OR
6	ENDSUBQUEST	B 200
END SUCB3TX1		

search as "partial results", and the output further refined by applying the logic of the text search to these partial results. A text profile is shown in Table VI.

STRUCTURE SEARCH

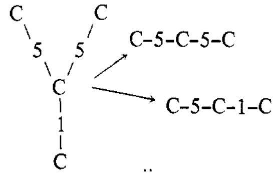
Lynch's group provided us with a frequency list, based on their sample of about 30 000 compounds from the CAS Registry, which gave the frequency of occurrence of atom- and bond-centered fragments at the most detailed level of description (called by us "bonded augmented atoms" and "bonded pairs", respectively). We calculated from this frequency list the frequency of the less detailed fragments higher up the hierarchy, by summing the frequencies of the daughter fragments below each parent. Once we had done this for the atom-centered, bond-centered, and four-atom-string fragments, we then chose an appropriate number of fragments to use in the screen, on the following principles: (a) every atom (or every bond) in the compound must be codable, so that the entire structure is in fact described by the code at some level of detail; (b) the screens should occur with equal frequency. The hierarchies of screens used by us, which differ slightly from Lynch's hierarchies, are shown in Table VII. We found it necessary also to introduce two new types of fragment.

One was the "decomposed" atom-centered fragment. This was needed because a



i

fragment (for example) actually contains within it a C–C–C fragment, and a searcher for the C–C–C fragment might not wish to exclude those compounds with the extra bond; on the other hand, he might wish to exclude them. Thus, atom-centered fragments with three or four nonhydrogen atoms attached to the central atom were broken down into decomposed atoms, e.g.,



ii

(type 5 bond = cyclic single bond, type 1 bond = acyclic single bond). Each of the three fragments shown would then be included higher up the hierarchy; thus ii would also give i and C3, and C-5-C-5-C and C-5-C-1-C would also give C-C-C and C2. However, the "decomposed" fragments were kept distinct from the corresponding nondecomposed ones, so that a searcher who wanted, for example, C-5-C-5-C positively not further substituted, could search for this without also retrieving C-5-C-5-C (decomposed).

The second extra type of fragment we introduced was the "other" fragment. Again the need for this is best explained by an example. Among the bonded augmented atoms chosen, on the basis of their frequency, for inclusion in the screen were:

(1.1.0)2C-(1)-P3(1.1.2)

(7.7.0)2C-(1)-P3(1.1.1)

(7.7.0)2C-(1)-P3(1.1.2)

Table VII. Hierarchies of Fragment Types

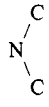
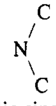
Name of fragment	Example	Structure of example
	(a) Bond-Centered Fragments	
Unbonded pairs	C-C	C–C (bond unspecified)
Simple pair	C-1-C	C–C (acyclic single bond)
Augmented pair	1C-1-C2	–C–C< (outer bonds unspecified)
Bonded pair	(2.0.0) 1C-1-C2 (1.2.0)	| =C–C= (acyclic bonds)
	(b) Atom-Centered Fragments	
Element	N	N
Coordinated atom	N2	N< (bonds unspecified)
Augmented atom	N2CC	 (bonds unspecified)
Bonded augmented atom	N2 1C 1C	 (acyclic single bonds)
	(c) Four-Atom-String Fragments	
Unbonded four-atom string	CCCX	C–C–C–X (bonds unspecified)
Bonded four-atom string	C1C1C2X	C–C–C=X (acyclic bonds)

Table VIII. Numbers of Different Types of Fragments in the Bit-Mask

Bond-centered fragments		617
Atom-centered fragments		
Elements	105	
"Decomposed" fragments	56	
Remainder of atom-centered fragments	141	
Total atom-centered fragments		302
Four-atom strings		115
Rings		122
Grand total		1156

Table IX. Ring Fragments[a]

Ring for-mula	Ring size											
	Fused rings						Monocycles					
	3	4	5	6	7	≥7	3	4	5	6	7	≥7
C only	1	2	3	4	5	6	61	62	63	64[b]	65	66
1 N	7	8	9	10	11	12	67	68	69	70	71	72
>1 N	13	14	15	16	17	18	73	74	75	76	77	78
1 O	19	20	21	22	23	24	79	80	81	82	83	84
>1 O	25	26	27	28	29	30	85	86	87	88	89	90
X	31	32	33	34	35	36	91	92	93	94	95	96
NO	37	38	39	40	41	42	97	98	99	100	101	102
NX	43	44	45	46	47	48	103	104	105	106	107	108
OX	49	50	51	52	53	54	109	110	111	112	113	114
NOX	55	56	57	58	59	60	115	116	117	118	119	120

[a] The numbers in the body of the table are the ring bit numbers. Atom types other than C, N, O are given the generic term X. For rings containing heteroatoms, only the heteroatoms are given; the other atoms in the rings are assumed to be carbon. Bit no. 0 is set for an "exception ring", that is, a fused-ring system that cannot be handled by the ring-analysis procedure. Bit no. 121 is set for a benzene monocycle, that is, a six-membered carbocyclic monocycle with aromatic bonds. [b] Bit no. 64 is *not* set for a benzene monocycle.

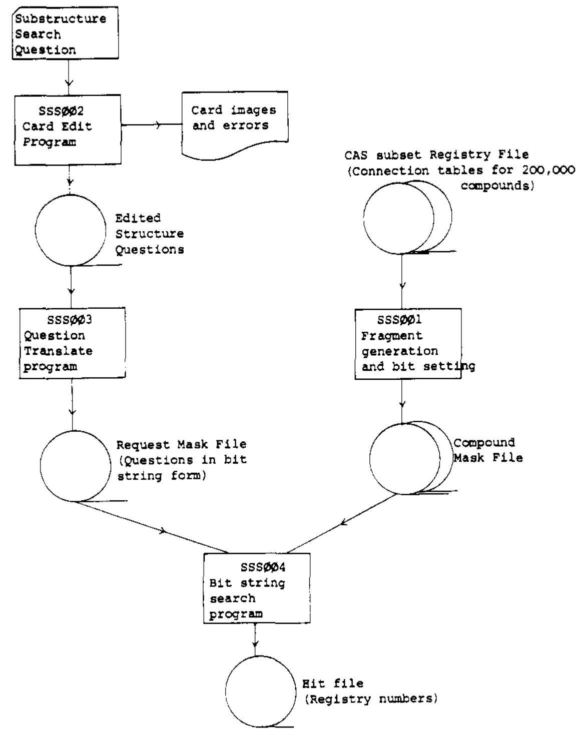

Substructure Search Question
SSSØØ2 Card Edit Program
Card images and errors
CAS subset Registry File (Connection tables for 200,000 compounds)
Edited Structure Questions
SSSØØ3 Question Translate program
SSSØØ1 Fragment generation and bit setting
Request Mask File (Questions in bit string form)
Compound Mask File
SSSØØ4 Bit string search program
Bit file (Registry numbers)

Figure 2. Substructure search system flowchart.

(The notation implies a single acyclic bond joining a carbon atom to a phosphorus atom. The phosphorus atom has three other nonhydrogen linkages; in the first example, these are two single and one double acyclic bonds. The carbon atom has two linkages and in the first example these are two acyclic single bonds).

There were other 2C-(1)-P3 fragments that occurred insufficiently frequently for inclusion in the screen. The problem then arises: how does one search for one of these rarer fragments? The fragment itself is not included, and the parent fragment 2C-(1)-P3 will be dominated by its three commoner daughters. So we introduced the "other" fragment—in this example "2C-(1)-P3(other)", which includes all those 2C-(1)-P3 fragments not otherwise coded at the bonded level. Thus a searcher looking for one of the rarer daughters of 2C-(1)-P3 would search for the "other" fragment and achieve reasonable precision, since the three commonest daughters would not be hit.

Table VIII lists the numbers of fragments used in the bit-screen.

In addition to the hierarchies of atom-centered, bond-centered, and four-atom fragments shown in Table VII, we also used ring fragments of the type used by Lynch's group.[8] These describe each ring (not each ring system) by its size and heteroatoms, distinguishing between monocycles and fused rings; the fragments are shown in Table IX. The program cannot deal with peri fusions or bridged systems; where these types of structure were found, the program set a special fragment (number zero), and also set all possible rings; e.g., if the ring system contained two nitrogen atoms, then all the fused nitrogen-containing ring fragments (no. 7–18) would be set, and so would the fused carbocycles (no. 1–6).

Figure 2 gives a system flowchart of the structure-text system. The most important program in the suite was the fragment generation and bit-screen setting program. This took the Registry File subset provided by CAS, of 209 211 compounds mentioned in Volume 76 of CA, as input; this file was in the Registry II format[3] as compact connection tables. The program was written in ICL System 4 Usercode, a very similar language to IBM BAL. The program (a) read a structure record and extracted data elements needed for the fragment generation—nodes, graph, bonds, and single-atom fragments; (b) set the element bits in the screen, from the nodes in the CAS record, translated the atom types (in "nodes") and the bond types (in "bonds") from CAS's hexadecimal codes into the codes we were to use, expanded CAS's compact connection tables, isolated bond-centered and four-atom fragments from the expanded connection table, matched them against the table of required fragments in the screen and set the appropriate bits "on" all the way up the hierarchy, isolated atom-centered fragments and similarly set all the appropriate bits "on", and finally pruned acyclic parts away and then set the appropriate ring-fragment bits "on", using Lynch's methods.[8]

To search the bit-screen data thus produced, we then wrote search profiles. For each of the queries in the chosen set—the same set that had been used for the nomenclature search—we wrote three search profiles. One was principally based on atom-centered fragments, one on bond-centered fragments, and one on four-atom-string fragments, though all three also contained ring fragments if the query involved cyclic compounds. The profiles were written quite mindlessly in an attempt to simulate the process that would occur if structure queries were input at a graphics terminal. Thus, for example, in a bond-centered profile, the technique was to write down the fragment based on every bond in the required substructure, described at the most detailed level, a bonded pair. If there was a variable or unknown bond on the substructure, then "or" logic would be required to allow for the alternatives at this point. The program which prepares the profile for search ("Question Translate" in Figure 2) then took the input fragments, coded at the most detailed level, and set the bits for the corresponding fragments that are actually present on the bit-screen, by a similar algorithm to that used in creating the data-base bit-screen itself.

Table X. Overall Number of Hits from Each Search

Type of structure search	Total no. of hits from structure search for all profiles (compounds)	Total no. of hits from structure-text search for all profiles (abstracts)
Nomenclature	22 374	5 802
Bond-centered	189 159	20 029
Atom-centered	236 795	29 076
Four-atom string	366 281	77 360
	Compounds	Abstracts
No. in Vol. 76	209 211	162 005
No. in $^1/_5$th sample		32 401

The search was then a straightforward bit-screen matching process, with only Boolean "and", "or", and "not" logic provided. The output from this search was in the form of Registry Numbers. These were then converted into the corresponding Abstract Numbers, and passed into the text search, which was identical with that which had been coupled with the nomenclature search, described earlier. Structure-text hits finally resulted; the queries were the same, and the text profiles were the same as those used for the nomenclature search, but there were three structure searches: atom-centered, bond-centered, and four-atom string.

OUTPUT

The hits from all four searches were then input to the UKCIS Housekeeping suite of programs.[17] With these programs, one can store the results from various searches of the same data base with the same queries using different variants of the profiles, and obtain comparative statistics of recall and precision between the different profile variants. Each hit is printed out for relevance judgment only once, regardless of how many variants of the profile may have retrieved it.

The hits from the different searches were output for assessment; the printout format is designed to allow one to write assessments upon it (R for relevant, X for irrelevant), and then to use the printout as a keypunching document for input of the assessments to the system.

For each hit, the printout gives only the CA abstract number. The abstract was then consulted in the printed CA issue; this and the notes of the user's query were then considered together and the relevance of the paper judged. The profile was not consulted, since the same information scientist (J.F.B.R.) wrote the profiles and performed the relevance judgments, and looking at the profiles at this stage might have caused bias.

The statistics programs in the Housekeeping suite use the nonparametric statistical method of Wilcoxon[18] for estimating significance levels.

RESULTS

Table X shows the total sizes of outputs obtained from the four types of search. The results are complicated by the fact that we had to restrict the nomenclature and text searches to a one-fifth sample of the file, owing to the core-store limitations of UKCIS's fairly small computer. The nomenclature-text searches, and the text part of the structure-text searches, were therefore carried out on a sample of the CAISF file, selected by choosing those abstracts whose Abstract Numbers ended with zero or five. In the structure-text searches it therefore follows that a substantial proportion of the structure hits were thrown away for lack of the corresponding text. This is unfortunate, but by performing the structure search on the full file we were able to gather useful timing figures.

The number of hits from the four-atom-string search was so large that the task of assessing them was judged too great. Nothing further was done with these hits.

The hits from the nomenclature-text, atom-centered fragment-text, and bond-centered fragment-text profiles were assessed and the results processed through the statistics programs; the final results are given in Table XI.

In theory, the relative recall of the structure searches ought to be 100%, since we tried to simulate an automated search of the fragments. In view of the lower values, we carried out a small-scale failure analysis. For a number of profiles where the recall of a structure-text search was less than 100%, we checked the hit printout and found a number of relevant papers hit by the nomenclature-text search but not by the structure search. We then consulted the user's original query, the profiles, and the abstracts in CA, to try to find the explanation for the failure to retrieve some documents by the structure search. Thirteen profiles were examined, including some atom-centered and some bond-centered profiles. In all cases, the recall failure could be attributed to human error, either in constructing the profile or in assessing the relevance of output. We therefore satisfied ourselves that our programs—both the search systems and the housekeeping system—were working correctly.

We also obtained, from the computer log, the times taken for the various different programs in the suite to run; we measured elapsed time, but the programs were not time-shared and most of them were C.P.U.-bound, so C.P.U. times were only a little bit shorter than elapsed times. The times taken are given in Table XII; we show not the times for individual programs, but aggregates for different parts of the system. The division of the incoming CAISF file into the text and nomenclature parts is necessary to provide a text file, and is therefore included under "preparation of general text search files", an overhead of 8.5 h on both the structure and the nomenclature methods.

The timing values need some interpretation in view of the fact that the nomenclature and text searches used the one-fifth

Table XI. Output Size, Precision, and Recall Results

	Nomenclature	Bond-centered	Std dev	Significance of difference
(a) Nomenclature Search vs. Bond-Centered Fragment Search				
Output size (av no. of documents/profile)	65.0	228.6	1.67	95% level
Precision (%)	59.6	41.7	2.70	99.5% level
Recall (%)	58.6	59.8	0.16	Not
(b) Nomenclature Search vs. Atom-Centered Fragment Search				
Output size (av no. of documents/profile)	65.0	332.6	3.06	99.5% level
Precision (%)	59.6	28.9	4.89	99.9% level
Recall (%)	58.6	61.6	0.12	Not
(c) Bond-Centered Fragment Search vs. Atom-Centered Fragment Search				
Output size (av no. of documents/profile)	228.6	332.6	1.32	Not
Precision (%)	41.7	28.9	2.17	97.5% level
Recall (%)	59.8	61.6	0.35	Not

Table XII. Computer Times Used

	Time used	Cost @ £48/h
Preparation of general text search files from CAISF	8 h 36 min	£412.80
Preparation of nomenclature search files from CAISF	9 h 35 min	£460.00
Nomenclature-text searching	6 h 46 min	£324.80
Preparation of bit-mask search file from CA Registry subset	11 h 27 min	£549.60
Bond-centered fragment-text searching	6 h 44 min	£323.20
Atom-centered fragment-text searching	7 h 6 min	£340.80
Four-atom fragment-text searching	6 h 46 min	£324.80

[a] These times refer to UKCIS's configuration of an ICL 4/50 computer with 128K of core and EDS60 60-megabyte discs. The current charge (from Oct 1, 1975) for outside users of this machine is £48/h.

sample file, while the structure searches used the full file. It took 11.5 h to produce the full bit-mask file, and about 7 h to do each structure-text search. The text part of the latter was on the sample file; one can estimate that if a full text file had been used, the time might have been 2.5 h longer. Thus the full one-volume structure-text system would have taken 11.5 h for file preparation and 9.5 h for search. By contrast, the nomenclature-text system took 9.5 h for file preparation and 6.75 h for search; but these figures are for the one-fifth sample file. We predict that a full one-volume search on the nomenclature-text system would have required more computer time than the corresponding structure-text search.

CONCLUSIONS

The results are finely balanced. Nomenclature searching gives the better results, since precision is better and recall about the same as in structure searching. However, structure searching was better able to handle the large files involved in searching 6 months' data at one pass on a relatively small computer and, although the results are not absolutely conclusive, it appears that structure searching would require less computer time for an equivalent amount of data. The nomenclature profiles were written as real-life profiles; we tried to write an optimum profile for the query. In the structure-text experiment, by contrast, we were testing the relative performance of atom-centered and bond-centered fragments, and we therefore wrote separate bond- and atom-centered profiles for each query. In real life, one would write a profile using atom-centered, bond-centered, and four-atom-string fragments together with ring fragments, all in the one profile as appropriate for the query. Presumably one could thereby achieve a superior performance to that available with any one type of fragment used alone. Since the difference in precision between the nomenclature-text and the bond-centered fragment-text profiles is not large, we surmise that optimum profiles, based on bond-centered fragments but enriched with other fragment types as required, would probably perform about as well as nomenclature-text profiles.

The quality of performance achievable by either nomenclature-text or structure-text searching, as measured by recall and precision values, seems to be in the range that commercial scientific information-retrieval services would regard as acceptable, in spite of the lack of an atom-by-atom search facility.

A further significant consideration is the ease of profile writing. Structure profiles are quite simple to write, given a structure diagram of the required substructure, and the process could easily be automated on a graphics terminal. Nomenclature profiles, on the other hand, are difficult to write and require a knowledge of chemistry and of CAS nomenclature; the process would be difficult to computerize. Lynch's belief that bond-centered rather than atom-centered fragments would give superior performance seems to be validated by our results.

A final consideration is the availability of the data bases. CAS does not publish subsets of the Registry file selected on a time basis, and thus one cannot normally obtain the set of connection tables corresponding to a particular volume. By contrast, they do publish *Chemical Abstracts Subject Index Alert* (CASIA), which contains the same data elements as CAISF, but arranged in document rather than alphabetic order, and appearing more frequently in smaller quantities. This data base is thus of more manageable size and can be concatenated up to whatever size is convenient for one's hardware. The regular availability of a convenient data base for nomenclature searching, and the nonavailability on a routine commercial basis of the connection-table data base, tips the scales back toward nomenclature searching again, and it is probable that if UKCIS offers a structure-text service commercially at any future time, it will use the nomenclature-text method.

ACKNOWLEDGMENT

We acknowledge the help and advice of Professor M. F. Lynch and his colleagues at the Postgraduate School of Librarianship and Information Science, University of Sheffield, U.K., and of Dr. G. G. Vander Stouw and his colleagues at the Chemical Information Science Research and Development Department, Chemical Abstracts Service, Columbus, Ohio. The programs described in this paper were written by the systems staff of UKCIS, the work being directed and to a great extent done by Mr. J. Powell. The work was supported financially by the British Library Research and Development Department (formerly the Office for Scientific and Technical Information), London, U.K.

LITERATURE CITED

(1) "Naming and Indexing of Chemical Substances during the Ninth Collective Period (1972-1976)", Section IV of the Index Guide, *Chem. Abstr.*, **76**, 32I-140I (1972).

(2) "Data Content Specifications for the CA Integrated Subject File in Standard Distribution Format", Chemical Abstracts Service, Columbus, Ohio, 1971.

(3) 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).

(4) G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153-159 (1973).

(5) J. E. Crowe, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. I. Non-Cyclic Fragments", *J. Chem. Soc. C*, 990-996 (1970).

(6) G. W. Adamson, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. II. Atom-Centred Fragments", *J. Chem. Soc. C*, 3702-3706 (1971).

(7) G. W. Adamson, D. L. Lambourne, and M. F. Lynch, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. III. Statistical Association of Fragment Incidence", *J. Chem. Soc., Perkin Trans. 1*, 2428-2433 (1972).

(8) G. W. Adamson, J. Cowell, M. F. Lynch, W. G. Town, and A. M. Yapp, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. IV. Cyclic Fragments", *J. Chem. Soc. Perkin Trans. 1*, 863-865 (1973).

(9) J. Powell, J. F. B. Rowland, and M. A. Veal, "Correlated Structure-Text Searching for Chemical Information", British Library Research Report No. 5292, 1976.

(10) I. C. McCracken, M. A. Veal, and S. Humphrey, "UKCIS Research Report No. 3. INFIRS—Inverted File Information Retrieval System", Office for Scientific and Technical Information Report No. 5166, 1973.

(11) J. E. Crowe, P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "Searching of Wiswesser Line Notations by Means of a Character-Matching Serial Search", *J. Chem. Doc.*, **13**, 85-92 (1973).

(12) P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "Evaluation of an SDI Service Based on the *Index Chemicus Registry System*", *J. Chem. Doc.*, **13**, 192-203 (1973).

(13) 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).

(14) "Chemical Abstracts Service Search Aids for the Ninth Collective Period (1972-1976)", Chemical Abstracts Service, Columbus, Ohio, 1974.

(15) D. C. Owsley, J. J. Bloomfield, A. J. Solodar, and E. Block, "CA Nomenclature", *Chem. Eng. News*, **53** (18), 3 (1975).

(16) M. Dub, "CA Nomenclature", *Chem. Eng. News*, **53** (18), 3 (1975).

(17) A. Robson, and J. S. Longman, "Automatic Aids to Profile Construction", Office for Scientific and Technical Information Report No. 5230, 1975.

(18) F. Wilcoxon, "Individual Comparisons by Ranking Methods", *Biometrics Bull.*, **1** (6), 80-83 (1945).

Comparison between CACon and CASIA Files for Development of New SDI Service in 1977

TILLY BAYARD and JEANNIE PERSOZ*

Association Francaise de Documentation Automatique en Chimie (AFDAC), 75116 Paris, France

Received December 21, 1976

A comparison between CACon and CASIA files is described, based on processing 40 profiles on both files. A new SDI service dealing with chemical structures and using CASIA tapes will be developed in 1977: chemical nomenclature, registry numbers, and molecular formulas will be searchable, thus allowing specific compounds and substructure searching at reasonable costs.

AFDAC is a part of CNIC, the French center licensed by Chemical Abstracts Service (CAS) since 1972. Our mission is to provide the French chemical community with computerized chemical information.[1]

At present we process 470 profiles on *Chemical Abstracts Condensates* (CACon) tapes, 70 profiles on *Polymer Science and Technology* (POST) tapes, and 20 profiles on *Chemical Industry Notes* (CIN) tapes.

Since 1974 we have processed on-line retrospective searches on CACon, and on other scientific, technical, and technico-economic files: BIOSIS, CAIN, COMPENDEX, INSPEC, METADEX, NTIS, PREDICASTS, etc.

AIM OF THE EXPERIMENT

In the field of chemistry, the problem of textual information searching has been given satisfactory solutions, both in Selective Dissemination of Information (SDI) and in retrospective search, with the CACon tapes.[2]

The situation is quite different for specific compounds or substructure searching. One of the most satisfactory solutions would be to process the CAS Registry File, with use of appropriate screens and possibility of atom-by-atom search. Several teams are working on this problem.[1,3-7] Two European industrial centers[5,6] provide their users with a retrospective substructure search service using the CAS Registry File. But, until now, such publicly available services did not exist.

Another way of searching the CAS Registry File is via chemical nomenclature.[8] Several American federal agencies use it for searching subfiles of the Registry File.[9-11] This solution seems especially interesting for SDI: even if services using the CAS Registry File with chemical codes become publicly available, updates of the file probably will not be very frequent because of their cost. Searching chemical nomenclature, available in the Chemical Abstracts Subject Index Alert file (CASIA), could be a good substitute for the update time interval.

Therefore, within a working group of the Union des Industries Chimiques (the French chemical manufacturers association which sponsored the foundation of AFDAC in 1970), it was decided, at the end of 1975 to compare the CASIA and CACon files for SDI, in terms of recall, relevancy, currency, and processing costs. This investigation also was expected to develop a thorough knowledge of CASIA, which starts being searchable on-line through the LOCKHEED/DIALOG system.

The development of a new operational service by AFDAC, using CASIA tapes, scheduled for January 1977, will result from this study.

DESCRIPTION OF THE FILES

File Coverage. The CACon and CASIA files include all the documents abstracted in the printed issues. They are organized sequentially, in the *Chemical Abstracts* (CA) number order. The CACon tapes are published weekly, while the CASIA tapes are published bimonthly and correspond roughly to an odd CACon tape plus an even CACon tape. However, index entries corresponding to some documents are published later in CASIA than in CACon. Thus, the corresponding CA index data in a CASIA tape does not include all the index entries for documents covered in the two corresponding CACon tapes, but it may include some for previous CACon issues. This study allowed us to compare the publication delay times of CACon and CASIA.

File Content.[12] The CACon file includes, for each citation, the CA number, title, author(s), bibliographic reference, organization, CA Section and subsection numbers, language, document type, issue keywords, and, for patents, the application country, number, classes, and priority dates. The title and issue keywords are written using uncontrolled vocabulary.

The CASIA file includes, for each citation, the CA number, CA section number, and index entries that will appear in the CA Chemical Substance, General Subject, and Molecular Formula Volume Indexes. These entries use controlled vocabulary at the heading level, though a thesaurus is not used. The chemical compounds are described according to the CAS index nomenclature rules, while general subjects are chosen from a list of predefined headings. The user can acquire a knowledge of this controlled vocabulary with the Index Guide, Volume Indexes, and CASIA Search Aids made available by CAS.

The indexing of documents in the CASIA file may be roughly described in the following way:

(a) Chemical substances are indexed by the CA preferred index name assigned according to the CAS substance index nomenclature and consisting of a heading parent, possibly followed by a substituent, and/or name modification, and/or qualifier, and/or stereo data element. The nomenclature is