An example of the PDP-10 program input and output is:

```
Type word for search specification
DNA
Found 175 references to word DNA
References Y/N?
N
Type word for search specification
RNA
Found 133 references to word RNA
References Y/N?
N
Reference conjunction
```

Conj No.	No. Abstracts	Words
1	20	DNA, RNA
2	133	RNA
3	175	DNA

Thus, in a matter of minutes of connect time and some 2 to 3 seconds of CPU time, the user identified 20 references that contain DNA and RNA. Even further refinement would have been possible if the user wished to narrow his question. It is important to note that only the CBAC digest number is available here as output, as opposed to the full abstract, title, author, etc.

This program has been very helpful for generating SDI profiles or retrospective search requests for the 360/370 programs because the user can quickly check the existence in the system of a variety of words and their possible synonyms (necessary because CBAC is free text and does not have rigidly limited vocabulary).

USER REACTION TO A BATCH SDI SEARCH SYSTEM

Preliminary studies indicate that the users agree on several points:

The PDP-10 search is impressive, it is useful in its own right and is a strong inducement to try the 360/370 SDI and retrospective searches. Its main flaw is no abstract.

The full abstract (available via the 360/370 programs) is the best part of CBAC. Users would be pleased if all of CA (and BA) contained full abstracts on tape. (Indeed some even said they would not consider using CA condensates or BA as they currently exist.)

Computer searching of data bases is valuable, but CBAC covers too narrow a range.

A number of users have expressed surprise to find journal articles retrieved that they had not seen previously.

CONCLUSIONS

The testing of CBAC is still in an experimental stage with a limited number of users. It is too early to determine if it will go into production status and, if so, whether it will need further refinements to achieve an optimal balance of the economic and informational needs of the NIH user community.

With the expansion of the size of the CBAC data base beginning in January 1972 (by about 40%), the range of coverage will obviously become greater and should please users. However, it remains to be seen if the intellectual value of increased coverage will offset the finanical cost of the larger data base.

Quite clearly when CA with complete abstracts are available in computer readable form, a searchable keyword section would have great economic value.

ACKNOWLEDGMENT

The authors thank Inez Gaffney and James Heilik of the National Science Library, National Research Council, for their assistance. We also thank Keatha K. Krueger, Victoria Ragen, and Billie Mackey for their aid in the Pilot CBAC Study.

LITERATURE CITED

(1) Feldmann, R. J., Heller, S. R., Shapiro, K. P., and Heller, R.S., "An Application of Interactive Computing—A Chemical Information System, "*J. Chem. Doc.*, to be published.

(2) Heller, S. R., "Profile Design Manual," DCRT/CIS, CBAC Literature Project.

Chemical Substructure Index (CSI)—A New Research Tool

CHARLES E. GRANITO* and MURRAY D. ROSENBERG
Institute for Scientific Information
325 Chestnut St., Philadelphia, Pa. 19106

Received June 9, 1971

In recent years, the extremely limited value of traditional indexes for substructure searching has become well recognized. As a result, considerable research effort has been financed for substructure search systems. However, with few exceptions, this research effort has been aimed at developing computer systems, even though most chemists do not have immediate access to computers. Manual substructure searches remain laborious or impossible to perform. The development of "desk-top" manually searchable indexes by permuting line notations was first reported in 1963.[1] These indexes permit manual substructure searching that is impossible or impractical with traditional manual indexes such as the subject or formula indexes to *Chemical Abstracts, Index Chemicus,* or *Beilstein.* Various versions of permuted line notation indexes are now used by many large organizations for internal files. However, access to the open literature for the chemist interested in finding new compounds that contain a particular substructure was unachievable prior to the introduction of the *Chemical Substructure Index* (CSI).

* To whom inquiries should be addressed.

Indexes of permuted notations constitute one application of the Wiswesser Line Notation (WLN) in industrial organizations. The Institute for Scientific Information (ISI) has introduced an index of this type for new compounds being reported in the journal literature—more than 150,000 per year. This new tool, called the *Chemical Substructure Index* (CSI), was created solely for substructure searches. The creation, advantages, and limitations of the CSI are discussed along with criteria used for selecting the more than 1,000,000 entries for the 1970 Annual. Also discussed are design features of the CSI that enable chemists unfamiliar with WLN to conduct substructure searches.

COMPOUND	WLN
Phenothiazine	T C666 BM ISJ
Phenoxazine	T C666 BM IOJ
Phenylacetic acids	QV1R
Phosphorothioate, O,O-dimethyl	OPS&O1&O1
Phthalazine	T66 CNNJ
Phthalimide	T56 BMVMJ

Figure 1. Phosphorothioates in FFS dictionary

		ABSTR	CPD
OPS&O1&OR BG DSWN1&1 *1	GNOPRSW	172130	18
OPS&O1&OR BSWN1&1 *1	NOPRSW		1
OPS&O1&OR C DSWN1&1 *1	NOPRSW		20
OPS&O1&OR CG DSWN1&1 *1	GNOPRSW		19
OPS&O1&OR CSWN1&1 *1	NOPRSW		2
OPS&O1&OR DSWM1 *1	MOPRSW		7
OPS&O1&O1 *4N4&SWR D	NOPRSW		15
OPS&O1&O1 *1Y&NY&&SWR D	NOPRSWY		14
OPS&O1&O1 *3N3&SWR D	NOPRSW		13
OPS&O1&O1 *2N2&SWR D	NOPRSW		12
OPS&O1&O1 *2N2&SWR CG D	GNOPRSW		21
OPS&O1&O1 *2N2&SWR BG D	GNOPRSW		22
OPS&O1&O1 *1Y&MSWR D	MOPRSWY		10
OPS&O1&O1 *2MSWR D	MOPRSW		8
OPS&O1&O1 *1Y&1N1Y&&SWR D	NOPRSWY		16
OPS&O1&O1 *5N5&SWR D	NOPRSW		17

Figure 2. Phosphorothioates entered in *CSI*

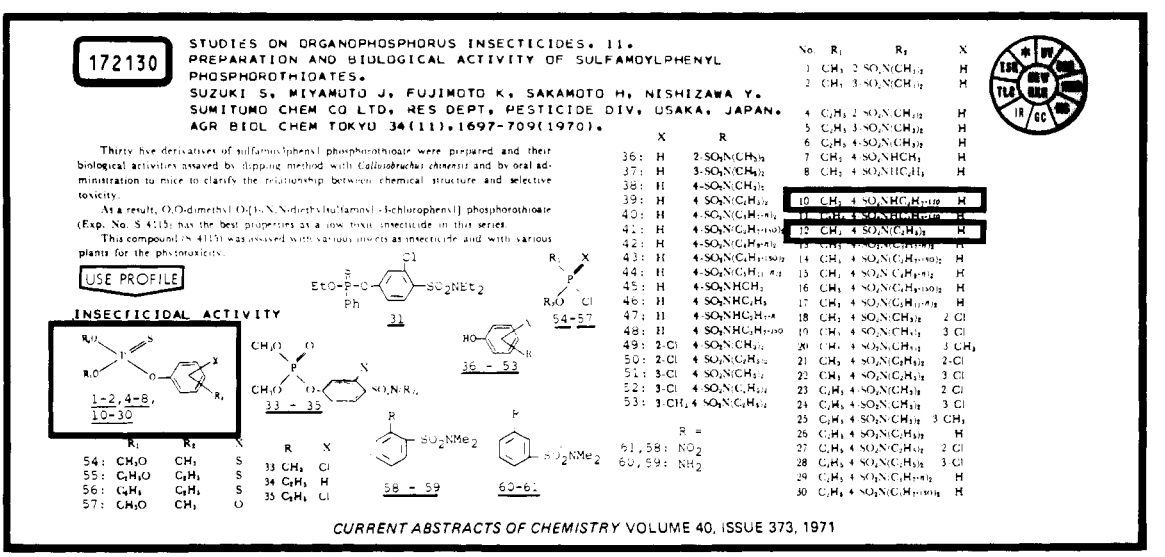
CHEMICAL SUBSTRUCTURE INDEX (CSI)

The *Chemical Substructure Index* is a monthly index of permuted Wiswesser Line Notations (WLNs) covering new compounds reported in the weekly issues of *Current Abstracts of Chemistry and Index Chemicus* (CAC&IC). The monthly issues are then cumulated annually. The first annual cumulation available covers the CAC issues for 1970. Present plans call for also producing 1967–1969 annuals. This will make substructure searching possible for over 800,000 unique compounds.

USE OF THE CHEMICAL SUBSTRUCTURE INDEX

Although it may take a full week's study of WLN to learn how to encode 90% of the organic compounds one will encounter in the literature and it may take as much as three months to become expert enough to encode compounds for ISI, any chemist can learn enough about Wiswesser Line Notations to use the *Chemical Substructure Index* in 20 minutes. To make CSI useful even to the chemist who has never seen Wiswesser symbols, ISI has introduced a number of aids. These are summarized in the User's Guide to CSI. First, a ready reference list of Wiswesser symbols is provided, explaining, in the order of their ranking, the 40 symbols used to make entries in CSI. Second, a dictionary of Frequently Found Substructures (FFS) is supplied. The following paragraphs describe the three methods of using CSI.

FFS Method. The FFS dictionary is an alphabetical list of common substructures or "parent" structures showing their corresponding WLN Notations. To use the FFS method, one simply looks up the generic name for a substructure in the FFS dictionary. The corresponding WLN is noted. Then the WLNs are scanned until the specific substructure is found (Figure 1). In this case, the individual interested in *O,O*-dimethylphosphorothioates in consulting the FFS dictionary would find the OPS&O1&O1 WLN symbols. By looking these up in the Index, he would find the new phosphorothioates which meet his request. Two of these are noted in Figure 2. Finally, the searcher


172130

STUDIES ON ORGANOPHOSPHORUS INSECTICIDES. 11. PREPARATION AND BIOLOGICAL ACTIVITY OF SULFAMOYLPHENYL PHOSPHOROTHIOATES.

SUZUKI S, MIYAMOTO J, FUJIMOTO K, SAKAMOTO H, NISHIZAWA Y. SUMITOMO CHEM CO LTD, RES DEPT, PESTICIDE DIV, OSAKA, JAPAN. AGR BIOL CHEM TOKYO 34(11),1697-709(1970).

Thirty five derivatives of sulfamoylphenyl phosphorothioate were prepared and their biological activities assayed by dipping method with *Callosobruchus chinensis* and by oral administration to mice to clarify the relationship between chemical structure and selective toxicity.

As a result, O,O-dimethyl O-[4-N,N-diethylsulfamoyl-3-chlorophenyl] phosphorothioate (Exp. No. S 4115) has the best properties as a low toxic insecticide in that series.

This compound (S 4115) was assayed with various insects as insecticide and with various plants for the phytotoxicity.

USE PROFILE

INSECTICIDAL ACTIVITY

CURRENT ABSTRACTS OF CHEMISTRY VOLUME 40, ISSUE 373, 1971

Figure 3. Phosphorothioates in *CAC*

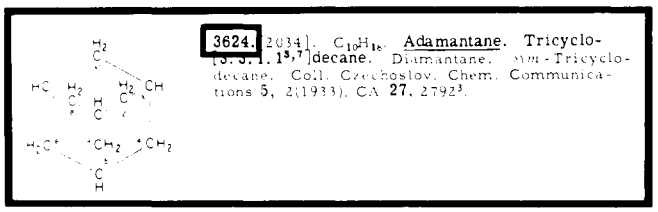


Figure 4. Adamantane entered in *The Ring Index*

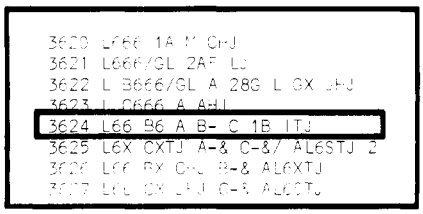


Figure 5. Adamantanes entered in *Wiswesser Line Notations Corresponding to Ring Index Structures*

would note the CAC&IC abstract number and compound number and consult CAC&IC (Figure 3).

Ring Index Method. The second method of using the CSI is the *Ring Index* method. To use this method, one must have a copy of *The Wiswesser Line Notations Corresponding to Ring Index Structures* (a document available from the Clearinghouse[2]). With this method, one consults *The Ring Index*[3] to find the ring system of interest. When the ring system is located, the *Ring Index* number is noted. This number is then found in *The Wiswesser Line Notations Corresponding to Ring Index Structures*, and the corresponding WLN is noted. Once the WLN is found, this is looked up in CSI. All new compounds containing the ring system noted would be found in one place within the CSI. Figure 4 shows the *Ring Index* entry for adamantane. If one were interested in new compounds containing this ring system, one would consult the *Ring Index* and find that adamantane has the *Ring Index* number 3624. By consulting the *Wiswesser Line Notations Corresponding to Ring Index Structures* (Figure 5), one would find that 3624 has the line notation L66 B6 A B- 1B ITJ. Using this to consult the CSI, one would find all new compounds containing the adamantane ring system located in one place. Those for Abstract 171543 are isolated in Figure 6. The corresponding structures are shown in the CAC&IC abstract (Figure 7).

WLN Method. The third method requires knowledge of the Wiswesser Line Notation. The index is then consulted without reference to the FFS dictionary or the *Ring Index*. Rather, the user looks up directly the WLN symbols corresponding to the structural feature(s) of interest.

PERMUTATION—A SPECIAL FEATURE OF CSI

As can be seen in Figures 1 and 4, the WLNs are permuted—i.e., each CSI entry was created by rotating the WLN to an appropriate index symbol.

Wiswesser Line Notations are on the average 20 charac-

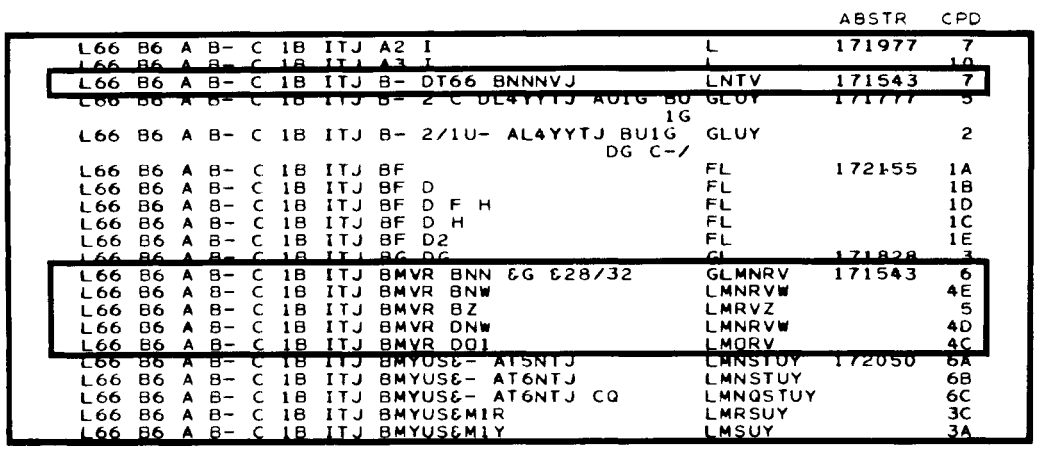


Figure 6. Adamantanes entered in *CSI*

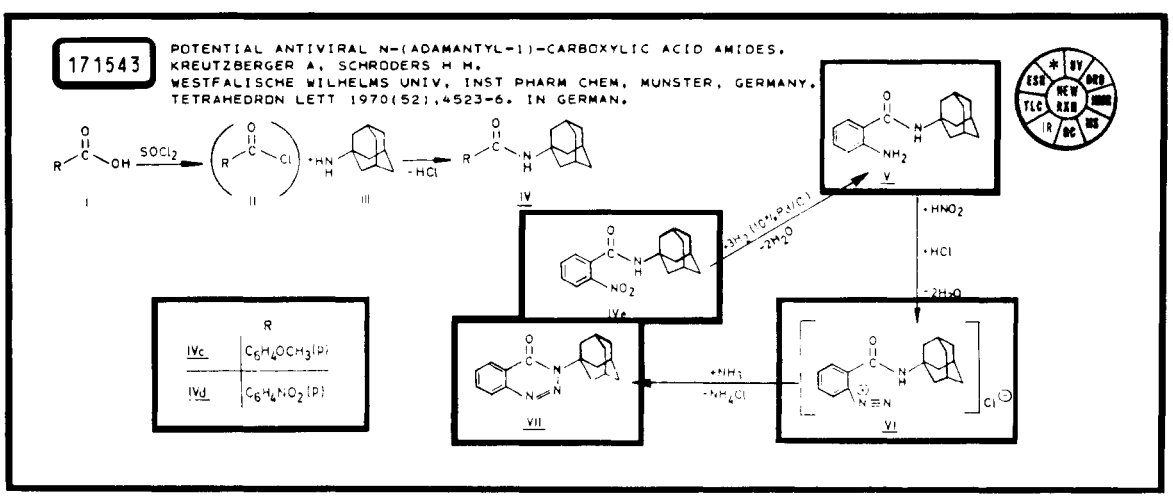


Figure 7. Adamantanes in *CAC*

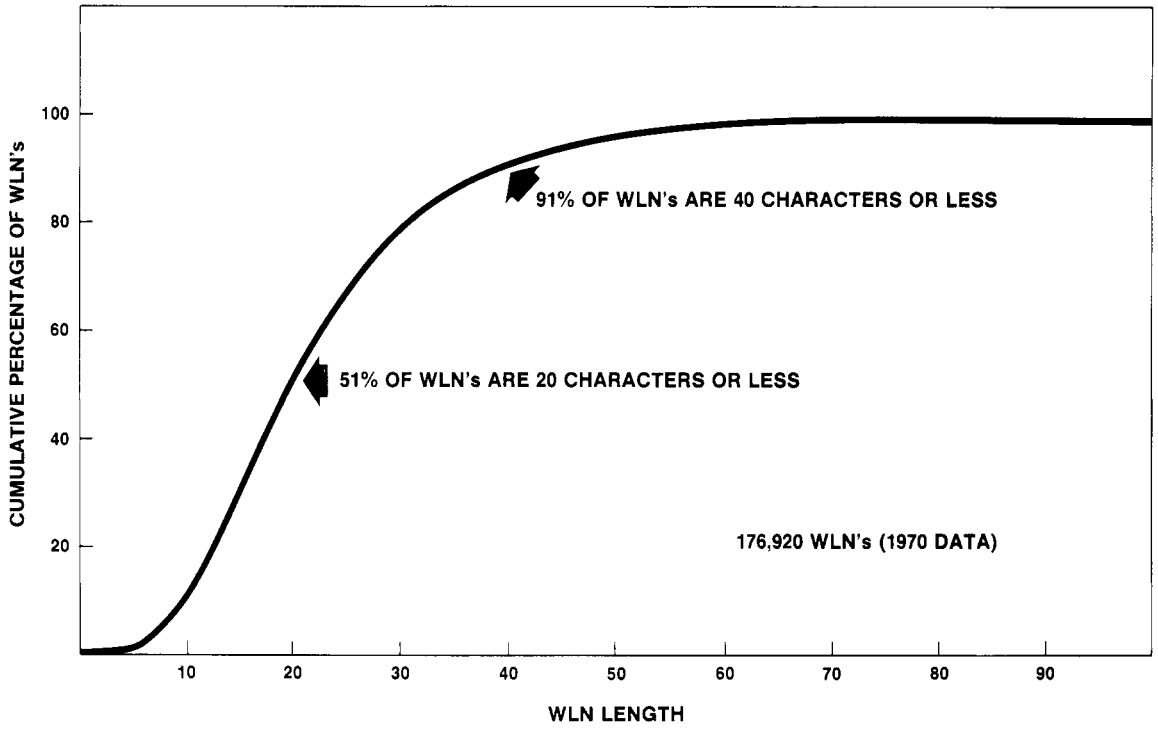


Figure 8. Number of characters in Wiswesser Line Notations

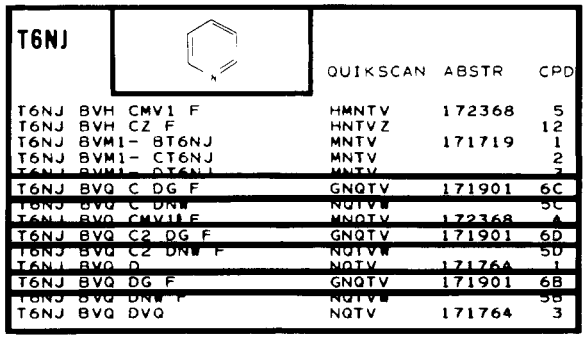


Figure 9. Pyridines entered in *CSI*

ters in length (Figure 8). Of these 20 characters, five to six are pertinent enough to justify an entry in a permuted index. Note that the entire WLN is always shown for each entry in the index. As a result, one can immediately locate compounds represented by a particular symbol initiating an entry. Additional fragments can be used to narrow the search instantly to a more specific area. This is illustrated in Figure 9. Pyridines are all located in the T section with all other heterocyclic compounds. The T6NJ (pyridine) symbols can be readily found. Furthermore, one can search for substituted pyridines (e.g., all chlorinated pyridines) by reading the symbols following the T6NJ's or by taking advantage of QUIKSCAN.

QUIKSCAN

QUIKSCAN is an alphabetized list of the WLN symbols which cause entries to be produced in CSI. However, it also contains X, Y, R, and H symbols (nonlocants). The latter are included in QUIKSCAN because they are useful secondary search terms. For example, if a searcher is interested in benzoic acids he would not want to start his search with R. About 40% of the new compounds contain an unfused benzene ring,[4] and R is therefore not indexed as a primary term. The searcher would first look up QVR and then VQ. The VQ section would contain a large number of entries. By including R in QUIKSCAN he can rapidly cover whole columns searching for VQ entries that contain R. To illustrate QUIKSCAN further, consider the search for all chlorinated pyridines noted in Figure 9. One could read all the symbols following the T6NJ's in search of those containing chlorine (G). However, it is far faster to read the QUIKSCAN column in search of G, particularly since QUIKSCAN is alphabetized.

Three compounds from CAC&IC Abstract 171901 are quickly spotted. Note that in QUIKSCAN a symbol can only occur once, regardless of how often it appears in the WLN. This is to conserve space and expedite scanning.

EXCLUSION RULES

To avoid large numbers of useless entries, certain WLN symbols (or symbols in a particular context) are excluded as main index entries. Figure 10 summarizes the excluded symbols. As a result of the exclusion rules, an average of slightly fewer than six entries per WLN was obtained for the approximately 176,000 WLNs processed in 1970. Of course, all symbols could be searched for (in any combination) by computer using the *Index Chemicus Registry System* tapes.[5] This backup is one benefit of having a machine readable record.

STRUCTURES

The CSI hand-drawn structural diagrams help in using this product. These structural diagrams are placed throughout the index according to frequency of occurrence. Within the computer program, a counter keeps track of the frequency of occurrence of symbols, and a list of the most frequently occurring sequences are supplied to ISI chemists

for a decision on what structures to provide. The inclusion of structures facilitates use of the Index.

PEPTIDES

A separate section of CSI contains all new peptides. Since a special one-letter code[6] is used for amino acids occurring in peptides, these entries are separated from the WLN section. The researcher interested in specific amino acid sequences will find this section especially useful.

SUBSTRUCTURE STATISTICS

Figure 11 shows the most frequent six or more character WLN sequences for 1970 (over 175,000 compounds). Not surprising is the fact that the steroid nucleus is the most common large fragment (6 or more WLN symbols) found in new compounds being reported in the literature.

This information should be useful to those interested in fragment codes or screens. Single symbol frequencies were reported earlier[4] and the 2 to 5 range is now being studied.

LIMITATIONS

Most of the substructure questions can be routinely answered through the use of CSI. However, there is a small percentage of substructure questions which cannot be easily handled with such a manual index. Included in this category are some questions requiring atom-by-atom searches—for example, all compounds containing a carbon atom connected to an oxygen atom which is two carbons removed from a nitrogen atom. Although this type of search could be performed with CSI, it would be quite laborious. This type of question is best handled with a computer.

Space limitations require that only an abstract and compound number be presented with each entry in CSI. Therefore, one must always return to *Current Abstracts of Chemistry and Index Chemicus* for additional information, such as the original journal citation. This is in contrast to a computer search where additional records such as journal citations can be printed out in answer to the search question.

SPECIAL NOTE FOR CSI SEARCHES

Because the WLN is linear, the symbols being searched may appear in two arrangements. For example, NW or WN for the nitro group. This should be kept in mind whenever a substructure search is being performed using CSI.

ADVANTAGES OF THE INDEX

CSI permits the chemist to do manual searches based on substructure alone. Since the data base is CAC&IC, the compounds being searched are all the new compounds being reported in the literature. Literally within minutes, a chemist can find all of the new compounds containing any specified substructure. He can, therefore, reduce the number of hours he must spend in the library and expedite his research projects. He can for the first time conduct

Locants and Numerals

All locants (letters preceded by a space) and numerals are excluded as primary index entries. They are also excluded from QUIKSCAN.

Special Characters

All special characters except hyphens are excluded as primary entries.

Hyphens

All hyphens are excluded as primary entries and QUIKSCAN entries *except* those used to initiate two-letter atomic symbols. For example, -Al- for aluminum. In the latter case, they create entries and thus bring together in the front of CSI all metal-containing compounds.

R and Y

R and Y have little or no primary indexing value and occur very frequently. They are, therefore, excluded as indexing symbols, but retained in QUIKSCAN.

X

X only creates an entry when used to represent a spiral point in a ring. However, all X's are retained in QUIKSCAN.

J

Since only specific compounds are included in CSI, J is excluded as a primary and QUIKSCAN entry.

H

H is always cited after the symbol to which it is attached. For example, VH for aldehydes. As a result, H has no primary indexing value and is excluded. However, when used to represent deuterium (H-2) or tritium (H-3), H is retained as a primary indexing symbol. Consequently, all deuterium and tritium compounds are brought together in CSI under H-2 and H-3 respectively.

T

When used to initiate a heterocyclic ring, T is used as a primary index symbol and included in QUIKSCAN. T is excluded from from both CSI and QUIKSCAN when used as a saturation mark.

Repeating Letters

Whenever a letter repeats itself in sequence, for example, GGG, only the first occurrence is used for creating an entry. In this way, extraneous entries are avoided. If the first symbol represents a locant, it is not considered. For example, a WLN containing the sequence NNNN (for an azide at the N position of a ring system) would receive an entry under the second N but not the first, third, or fourth.

Special Cases

Metals

Metals are indexed only once on the first hyphen. The two letters and second hyphen are excluded.

Multi-cyclic points and Bridge atoms

All multi-cyclic points and bridge atoms are excluded as indexed entries.

Single Ring Atoms

Potential entries *starting* with the following WLN symbols are excluded: M 1, MJ, MT, N 1, NJ, NT, ... (where 1 = any letter)

These ring segments have little, if any, value as *primary* search terms and are, therefore, excluded. However, the initial symbols are retained in *QUIKSCAN*.

NOTE: The rules noted above do not apply to the initial symbol of any WLN. Therefore, one can easily use CSI for specific compound searches as well as substructure searches.

Figure 10. Exclusion rules

1970 CSI Data—176,000 WLNs

WLN Symbols	Freq.	WLN Symbols	Freq.	WLN Symbols	Freq.
L E5 B6	4990	N HNJ	1157	NUNR D	754
T6OTJ	4376	O- BT6	1148	L6UTJ A	749
-SI-1&	3246	OV1 DO	1094	SWR D&	745
T56 BN	2870	O1 EO1	1078	NR CNW	742
L6TJ A	2772	V1 DOV	1071	T56 BOJ	738
T5OTJ	2767	L50J O-	1048	WNR CN	733
T6NJ B	2542	OV1 EO	1040	L66J C	714
L66J B	2363	T6N CN	1029	L5OJ A	713
UTJ A E	2336	N CN E	1024	T5SJ B	710
Q DQ E	2322	T5NN D	1021	WR D&	698
T66 BN	2040	VTJ A	1019	PO&O2&	685
O1 DO1	1890	UNMR B	1015	T5NNJ A	670
N CNJ	1781	L3TJ A	986	T56 BM	660
T6NTJ A	1772	MR BNW	982	-FE--	655
OTJ CQ	1771	N DNJ	978	N DNTJ	651
T C666 B	1763	T5OJ B	967	O EVJ	648
OTJ B1	1758	OSWR D	964	O GOTJ	632
N DOTJ	1638	T B656	963	T66 BM	631
T56 BN	1631	V1 EOV	959	VNVJ C	630
T6N DO	1578	N FN H	959	T6N DNTJ	630
N DN F	1529	T5O CO	955	T3OTJ	607
T56 BM	1493	OTJ CO	920	NUNR B	601
Q EQ F	1486	T B666	898	OV1 F1	597
N ENJ	1463	O-SI-1&	881	V BUTJ	592
T66 BO	1446	NMR BNW	878	T55 BO D	591
T6N CNF	1416	T66 BN	857	G DG E	590
OTJ BO	1366	PR&R&R	831	C-14 &	588
O COTJ	1302	O1 IO1	831	T6NJ C	587
NW DNW	1229	T6NVMV	824	UTJ B	581
T56 BV	1212	OPQO&O	813	M DNJ	581
OTJ B	1198	T56 BO	796	NTJ Al	580
T66 BO	1187	OTJ C	780	O DO G	576
T66 BV	1163	T5NTJ A	763	T5N CS	567
				NTJ AV	565

Figure 11. One-hundred fragments (6 or more symbols) appearing most frequently in 1970 *CSI*

searches which heretofore were impossible. The advantages of using CSI are both short- and long-range. The short-range advantages are demonstrably economic and include reduced product development costs, shorter research time, and avoidance of duplicative research. The long-range advantages, while less obvious, are hardly less important. The history of research and development shows that chemical discoveries have led to the introduction of new industry or great expansion of existing technologies. The use of CSI by those engaged in research and development tasks should aid in making similar advances.

LITERATURE CITED

(1) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line-Notations," *J. Chem. Doc.* **4**, 56-60 (1964).

(2) PB 180 901, "Wiswesser Line Notations Corresponding to Ring Index Structures," Chemical Abstracts Service, distributed by Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. 22151.

(3) "The Ring Index, Second Edition," Chemical Abstracts Service, Columbus, Ohio 43210.

(4) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.* **11**, 106-110 (1971).

(5) Garfield, E., Revesz, G. R., Granito, C. E., Dorr, H. A., Calderon, M. M., Warner, A. W., "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemica Retrieval," *J. Chem. Doc.* **10**, 54-8 (1970).

(6) Revesz, G. S., Granito, C. E., and Garfield, E., "One-Letter Notation for Calculating Molecular Formulas and Searching Long-Chain Peptides in the Index Chemicus Registry System," *J. Chem. Doc.* **10**, 212-16 (1970).

Syntactical Proximity—Partial Syntactical Analysis of Natural Language Data Records

TAKEO YAMAMOTO* and SHIZUO FUJIWARA

Department of Chemistry, Faculty of Science,
The University of Tokyo, Hongo, Tokyo 113, Japan

Received May 13, 1971

A definition is given for the syntactical proximity of a string in a natural language data record. It uses two dictionaries for the "delimiters" and a limit of length n_{prox} in the search for the proximity. It is shown that, by giving suitable entries for the dictionaries, several kinds of partial syntactical analyses may be performed by a relatively simple operation.

Information retrieval operations using natural language data bases such as CAS data tapes are mostly done by the term-match method. The query provides the system with a dictionary of meaningful terms—words, phrases and/or fragments thereof. A record is retrieved whenever it contains the required combination of the terms. To retrieve records containing information about iron complexes, for example, one provides a dictionary consisting of two sets of terms ("parameters")—one with terms such as "ferrous ", "ferric " and "iron", and the other, with terms

* To whom inquiries should be sent.