The Searching of Wiswesser Line Notations by means of a Character-Matching Serial Search

J. E. CROWE,* PETER LEGGATE, B. N. ROSSITER** and J. F. B. ROWLAND*‡
University of Oxford, Faculties of Physical and Biological Sciences,
Experimental Information Unit, 7 Keble Road, Oxford OX1 3QL, U.K.

Received March 8, 1973

The *Index Chemicus Registry System* (ICRS), the machine-readable version of *Current Abstracts of Chemistry and Index Chemicus*, has been used to provide a computer-based SDI service to 183 research workers in industrial and university laboratories. The techniques of search strategy design for substructure searches of the WLN file are described in detail.

In the last two decades, much effort has been devoted to the creation of machine-readable chemical structure files.[1,2] Much of the early work in this field was undertaken by the larger chemical and pharmaceutical companies[3-5] and was concerned with proprietary files of compounds for use within the company. More recently, several machine-readable files of compounds in the published literature have been created.[6-8] One such file is the *Index Chemicus Registry System* (ICRS) which is produced by the Institute for Scientific Information (ISI).[7]

ICRS uses the Wiswesser Line Notation (WLN)[9,10] for representing the structures of chemical compounds in a machine-readable form. This notation represents the molecule by a string of alphanumeric characters, and it is possible to search a file of such notations by means of a character-matching program similar to those used for searching natural-language text records.

The ICRS magnetic tape files are issued monthly; each tape contains the data from one month's issues of the weekly abstracts journal *Current Abstracts of Chemistry and Index Chemicus* (CAC&IC). These files can be used to provide a monthly selective-dissemination-of-information (SDI) service. For each document abstracted in CAC&IC, the tape file contains the index records shown in Table I.

Table I. Data Elements Present in ICRS for Each Document

1. Wiswesser line notations of all new compounds
2. Molecular formulae of all new compounds
3. Title (in English)[a]
4. Author(s)[a]
5. Address of author(s)
6. Journal citation[a]
7. Language of original article (if other than English)
8. Subject index words—natural language keywords assigned by indexers[a]
9. "Use profile," indicating proven uses of compounds, for example, ANTIMICROBIAL ACTIVITY or DYE
10. "Instrument data alert," indicating the analytical techniques whose use is reported in the paper.

[a] Present for all documents.

A WLN and a molecular formula is given for every new compound in the paper; the average number of compounds indexed on the WLN file is 10 per paper. (ISI assume that a compound is *new* if a paper cites no reference to a previous preparation.)

From 1969 to 1971, the Experimental Information Unit operated an experimental SDI service based on the ICRS tapes. The objectives of this experiment were:

A. To study techniques of search strategy design for substructure retrieval by string searching of WLN, and to define the scope and limitations of WLN string searches.

B. To use both the structure and the bibliographic files of ICRS as a basis for an experimental SDI service for academic and industrial research workers

C. To assess the nature and extent of user demand for services providing substructure retrieval facilities

D. To evaluate the service in terms both of quantitative measures of effectiveness and of user reaction, to identify the major deficiencies of the service and, if possible, to suggest methods of improving performance

The present paper gives a brief description of the establishment of an experimental SDI service (objective B) and discusses in detail the techniques of substructure search using a WLN file (objective A). A later paper[11] will describe the evaluation of this service (objectives C and D). A description of the complete experiment is available in report form.[12]

THE EXPERIMENTAL SERVICE

The Search Programs. ISI supply a suite of programs (RADIICAL) to all subscribers to the ICRS tapes. The use of these programs for searching WLN records has been described by Granito *et al.*[13] These programs were not released in the United Kingdom in their complete form until November 1970 and were not therefore available when the present project started. It was therefore decided to use a modified version of the serial search programs which had been developed by the United Kingdom Chemical Information Service (UKCIS). These programs were written in KDF9 Usercode and had been designed for use in searching Chemical Abstracts Service (CAS) natural-language data bases. The modified serial search provided special facilities for searching WLN records and also allowed for coordinate

*Present address: United Kingdom Chemical Information Service, University of Nottingham, Nottingham NG7 2RD, U.K.

**Present address: Computing Laboratory, University of Newcastle upon Tyne, Newcastle upon Tyne, U.K.

‡ To whom correspondence should be addressed.

linkages between WLN and text terms (e.g., L E5 B666 *and* SPECTR). The logical facilities provided by these programs, as modified for the WLN search, are shown in Table II. Programs were implemented on the Oxford University KDF9 computer. (Program modifications were carried out by A. K. Kent and I. C. McCracken in collaboration with the present authors.)

The User Population. Users for an experimental service, based on the ICRS tapes, were recruited by sending out a circular letter to organic and pharmaceutical chemists in selected university departments and to information officers in a number of large industrial organizations. To enable research workers to decide whether the subject coverage of the ICRS data base was appropriate to their research interests, the journal coverage and article selection criteria of ICRS were explained to them, both in circulars sent to potential participants and at departmental seminars where these were held.

Table II. Logical Facilities Provided by the Modified UKCIS Serial Search Programs

Logical Facility	Word Terms	WLN Terms
Search terms	Two character minimum	One character terms allowed
Truncation	Front and rear-end truncation	Same
and	Boolean *and*	Same
or	Boolean *or*	Same
not	Boolean *not*	Same
with[a]	A Modified Form of *and* Logic Which Allows One to Stipulate:	
	That the two search terms must co-occur within the same sentence, or that one term must follow the other within a specified number of words	That the two terms (i.e. WLN fragments) must be in the same *notation*, or that one term must follow the other within a specified number of *characters*. There are two limitations on the use of *with* logic: it can only be used at a coordination level of 2 (it is not possible to specify A *with* B *with* C), and it cannot be used if both terms are the same (A *with* A)
ignore	A Modified Form of *not* Logic Which Offers the Same Additional Facilities as *with* Logic	
internal truncation	Not available	Implemented to circumvent the limitations of *with* logic mentioned above. Allows searches such as: A *followed by* A *followed by* B *followed by* C. For each *followed by* instruction, a maximum number of intervening characters is stipulated. Thus: A(p)B(q)C .... stipulates: A followed by B within p characters, followed by C within q characters . . .

[a] Combinations of *with* logic and internal truncation were used to specify the various forms of *followed by* logic used in the examples appearing in the text.

All research workers who volunteered to take part in the experiment were interviewed (by B.N.R. or J.F.B.R.) on at least two occasions to define their current awareness requirements and to negotiate suitable search strategies. Particular attention was paid at the first interview in obtaining an accurate and detailed statement of requirements. The second interview took place after the user had received two or three months' output from his profile. The profile was often amended substantially during the second interview on the basis of the results obtained in these initial searches.

A pilot group of 36 users was recruited in November 1969, and this group was expanded to a full population of 183 users between April and October 1970. The population comprised 127 academic research workers from 31 different university departments and 53 industrial research workers from seven establishments. Three users from a government research laboratory were also recruited. Of these 183 users, 102 took part in the subsequent evaluation experiment.[11]

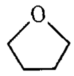
User Requirements. It was our intention to offer participants a service which would satisfy as many of their current awareness needs as were appropriate to the subject coverage of the data base. We wished to avoid 'model queries' designed to test the capabilities of a structure search system but not representative of the research workers' main interests. Participants were not asked to restrict themselves to concepts of an exclusively structural nature, though in practice such concepts were a major feature of almost all requirements.

As a result of this emphasis on broad current awareness needs, user requirements were often complicated and involved searches for a number of different structural fragments rather than for a single fragment. This was especially true of university staff members who had several research students working on different, though related, research projects. Examples of user requirements are:

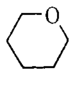
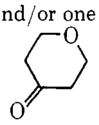
Example 1. A user with interests in several distinct types of structure: 'Derivatives of cyclobutenes and cyclobutadienes, especially biphenylene and its heterocyclic analogues.
Preparation of quinones and the study of their addition reactions.
Chemistry of triazoles.
Benzyne and the arynes.'

Example 2. A typical, fairly straightforward, query: 'Compounds containing 8-, 9-, 10- or 11-membered rings with one or more double or triple bonds on the ring. The rings may be carbocyclic or heterocyclic, and the compounds may contain either isolated rings or fused rings.'

Example 3. A user with a very broad requirement: 'Monosaccharides, oligosaccharides and their derivatives:



(with optional carbonyl groups and/or one double bond)

 (with optional carbonyl groups and/or one or two double bonds, but excluding )

with any single bonded substituents.
Also cyclic acetals and ketals of these carbohydrates.'

Example 4. A user with a primary interest in a nonstructural concept, but with some structural aspects to the requirements: 'Generally, mass spectrometry (m.s.) of organic compounds. Specific aspects of the interest were: pyrolysis-gas chromatography-mass spectrometry of organic compounds (not polymers); photoionization m.s. of organic compounds; m.s. of ergot alkaloids; m.s. of sesqui- and di-terpenes; substituent effects on electron-impact-induced fragmentation mechanisms; m.s. of ^{13}C-labelled compounds; the use of "metastable spectra" in structure determination by m.s.; m.s. of compounds containing the hexahydrophenanthrene structure.'

Search Profiles. The WLN structure file is the most important feature of the ICRS data base, and an interest in chemical structures was a dominant feature of the current awareness requirements of most recipients of the SDI service. In constructing search profiles our attention was, therefore, largely concentrated on searches of the WLN file. However, title words and other data elements listed in Table I were also used in the search profiles, when appropriate.

The liaison scientist encouraged the use of broad profile strategies designed to provide maximum recall, at least in the initial profile. If the first two or three monthly outputs produced by this broad profile were unacceptably large, amendments were made at the second interview. This emphasis was accepted by most users and, even after amendment, the majority of profiles were directed to the achievement of high recall rather than high precision. This is reflected in the very large outputs received by some users (nine users had an average monthly output of over 100 references, though the average for all users was 42 references per month).

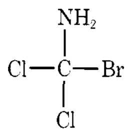
GENERAL PRINCIPLES OF WLN SEARCHING

On the basis of experience in profile construction gained in the experiment, general principles relating to the design of substructure search strategies for a WLN file were established. These principles of search strategy design are applicable to any character-matching search of WLN and are not dependent on the particular type of search program used. On the other hand, the method of coding these strategies in a search profile is dictated by the search program which is being used. Though the serial search programs were not originally intended for use in searching WLN records, the modified version of these programs provided, with a few minor exceptions, all of the essential logical facilities required for such searches. However, the profile coding was often cumbersome. The advantages to be gained by the use of custom-designed programs and, in particular, the use of bit-screen techniques,[13] are the simplification of profile coding and the reduction of search costs.

The main advantage of using WLN for a structure file is its compact character, which minimizes the size of the structure records that have to be stored and scanned. To achieve this compactness, two principles are paramount in WLN coding:

> The notation takes only one path through a compound's structure, and the starting point and routes through branch points are chosen according to well-defined rules. These rules can give completely different notations for structures which are closely related.
>
> Certain structural features are omitted from the notation. Their presence in a structure is inferred when a manual decode is made.

Even with the most sophisticated search programs, these two principles pose problems in a character-matching string search because many of the atom-to-atom connections are not explicit. For example, the structure

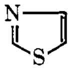
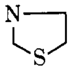
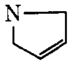


is coded as ZXGGE. Although the fragments ZX and XG indicate that NH_2—C and C—Cl bonds are present, the symbol, E, is separated from the symbol, X, and the C—Br bond is not readily apparent in a machine search.

In principle, it is easier to search effectively for a cyclic structure than for an acyclic structure. This is because the path of the notation through a cyclic structure always starts on one ring atom and passes through all the remaining ring atoms before entering the substituents. Thus the coding rules exclude the many possible notation paths which would involve citing substituents before citing the ring. Another useful feature of the notation in searching for ring systems, which has no equivalent in acyclic structures, is the use of locants to define the position on the ring of both hetero-atoms and substituents. However, the separation of a ring from its substituents in the notation creates severe difficulties in searching for fragments which may be located partly in a substituent and partly in a ring. Substructure search strategies will be described below.

Searches for Cyclic Substructures. *General Comments.* If a research worker requires information on compounds which contain a specified ring system, it is imperative to determine whether the ring system is still of interest when other rings are fused to it. The search is relatively straightforward if the presence of additional fused rings can be excluded, and the specified ring system is of interest only when *isolated* from other rings. This is because the notation path is determined on a hierarchical basis by (a) the position and type of ring fusion, (b) the ring sizes, (c) the positions and nature of hetero-atoms, and (d) the position of unsaturation. If an *isolated* ring system is specified, (a) and (b) cannot vary and only variations in (c) and (d) need be considered. In the simplest case of all, (c) and (d) are also specified and structural variations are confined to substituents on the ring system. The required ring fragment can then be retrieved by means of a single WLN search term. For example, the term T6NJ will retrieve all substituted pyridines. However, if the ring system is also of interest when other rings are fused to it, (a) and (b) may also vary and this results in an enormous increase in the number of possible notation paths. Consequently, this type of search poses many problems owing to the large number and variety of possible notations.

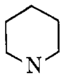
The Ring-Frequency List. A useful aid to designing search strategies was a list of all ring notations appearing on the ICRS tapes during a 12-month period, together with their frequency of occurrence. This frequency list was generated from the ICRS tapes by a modified version of a UKCIS program designed to count the frequency of title words. The program was modified by redefining a *word* as any string of characters beginning with L or T and ending with J (excluding J). (An underline is used to denote a leading or trailing space in a WLN search term. Thus [space]J is represented as _J.) It must be emphasized that this list does not contain complete notations, but only those stretches of notations corresponding to a ring system. An example of the entries on this list is:

WLN·	Frequency (Documents/ year)	Structure
T5N CSJ	21	
T5N CSTJ	5	
T5N CUTJ	5	

(The structural diagrams do not, of course, appear in the frequency list.)

Searches for Isolated Ring Systems Which are Not Fused to Other Rings. In searching for isolated ring structures we made use of two general types of search strategy which we termed *fragmented* and *unfragmented.*

Unfragmented Strategies. In an unfragmented strategy a search is made for the complete notations for all possible rings that are relevant to the user's interests. For example, a user may be interested in compounds containing the following carbon-nitrogen skeleton:



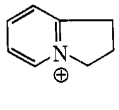
A user whose interests are restricted to pyridines and pyridinium salts will be satisfied by a search for T6NJ *or* T6KJ. If pyridones are also of interest, the number of possible notations is increased since ring carbonyl groups are coded as if they were hetero-atoms. Thus, T6NVJ, T6N DVJ, T6N CVJ, T6N CV EVJ, T6VNVJ, etc. become possibilities. If completely saturated (piperidine) structures are also required, the number of possible notations is large even for this apparently simple requirement.

A ring-frequency list can be used to identify all ring notations relevant to a requirement which have actually appeared in the data base during a 12-month period. An unfragmented strategy is often feasible if we accept the compromise of searching only for those notations which are shown by the ring-frequency list to occur in the ICRS file in at least two documents a year.

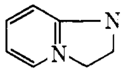
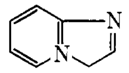
A variation of the unfragmented strategy can be used if all the possible notations have a common characteristic consisting of a single, unbroken stretch of notation. For example, there are many notations for steroid rings, corresponding to the presence of ring carbonyl groups and double bonds in different positions. A search for the notation string L E5 B666 will, however, find all steroids with the normal cyclopentanophenanthrene nucleus, and a relatively small number of other notations will cover nor- and homo-steroids and other variants. This type of search will be useful only if the characteristic is rarely found in other notations. It is not suitable in a search for a rare group if a much commoner group has a very similar notation—a 'needle in a haystack' search. An example would be a search for benzimidazoles (T56 BN DNJ and variants such as T56 BN DN FV IVJ). T56 BN DN would not be a suitable search term as this would also find papers on the more frequently occurring purines (T56 BN DN FN HNJ, etc.).

The advantages of unfragmented searches of the type described above are that they give good precision and relatively short search times since the use of complex logic is avoided. Their main disadvantages are that references are missed because search terms are very specific and that large numbers of search terms are often required.

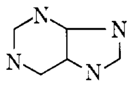
A related approach is to search for an appropriate common fragment of relevant notations and to eliminate notations that are not required by the use of *ignore* (restricted *not*) logic. A ring-frequency list is an invaluable aid in this type of search as it can be used to identify the notations that are not required but which would be retrieved by one of the main terms. In our view, it is usually not worthwhile to exclude a notation unless it occurred in at least two documents per year. An example of this type of strategy would be a search for indolizines and related structures in which the saturated ring carries one or more *exo* double bonds (including C=O) or spiro carbon atoms. It was decided that it would be easier to *ignore* the unwanted structures rather than to write down notations for all the acceptable variants.

Main terms	T56 AK T56 AN	
Ignore terms (within 6 characters)	_DM _DN _DK _DO _DS	
	_CN _CM _CK _CVN _CYN _AKB _AKM	
(and nine other fragments)		

The smallest possible WLN fragments are used in the *ignore* list so that each term will exclude as many notations as possible. For example, the three-character term _DN will exclude both T56 AK DNT&J and T56 AK DN BHJ, representing

 and 

Fragmented Strategies. The alternative *fragmented* strategy is used if there are a large number of possible notations for the required substructures, but all of them have two or more characteristic stretches of notation with unknown symbols intervening. An example is the purine nucleus:



The possibility of different degrees of saturation, of substituents on the nitrogen atoms, and of carbonyl groups, external double bonds, and spiro functions can lead to a great number of notations for this nucleus.

The 18 notations which occur in two or more documents per year in the frequency list are:

T56 BM DN FN HNJ
T56 BM DN FNVNVJ
T56 BM DN FVM INJ
T56 BM DN FVMVNJ
T56 BM DN FVN INJ
T56 BN DM FN HNJ
T56 BN DN FMVMVJ
T56 BN DN FMYMVJ
T56 BN DN FN HKJ
T56 BN DN FNVMVJ
T56 BN DN FNVNVJ
T56 BN DN FVM INJ
T56 BN DN FVN INJ
T56 BN DN BYM INJ
T56 BN DN FYN INJ
T56 BN DN FN HN CHJ
T56 BN DN FN HN IHJ
T56 BN DN FN HNJ

An appropriate fragmented search for this substructure might be:

T56 BM DN
or T56 BM DM
or T56 BN DN
or T56 BN DM
followed by _FN *or* _FM *or* _FVM *or* _FVN *or* _ FYM
followed by _VN *or* _VM *or* _HN *or*_HM *or* _IM *or*_ IN *or* YM

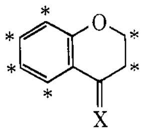
In this particular case, the fragmented strategy requires more individual search terms than a simple unfragmented search for the 18 notations appearing in at least two documents a year, but the profile may retrieve other relevant ring systems not predicted in advance.

A search for *all* theoretically possible purine notations, not just the frequent ones, would require the following fragmented search:

T56
followed by _BN *or* _BM *or*. _BK
followed by VN *or* VM *or* VK *or* XN *or* XM *or* XK *or* YN *or* YM *or* YK *or* _DN *or* _DM *or* _DK
followed by _FVN *or* _FVM *or* _FVK *or* _FXN *or* _FXM *or* _FXK *or* _FYN *or* _FYM *or* _FYK *or* _FN *or* _FM *or* _FK
followed by VN *or* VM *or* VK *or* XN *or* XM *or* XK *or* YN *or* YM *or* YK *or* _HN *or* _HM *or* _HK *or* _IN *or* _IM *or* _IK

The combination of *with* logic and internal truncation which is used to simulate the required logical relationships would involve 200 search terms and represents the upper practical limit in terms of profile size. The unfragmented approach exemplified by a search for steroids would not work well in this case. A search for T56 BN without further restrictions would also retrieve compounds with only one, two, or three nitrogen hetero-atoms.

A requirement for azaflavines and azachromones provides a further example of a fragmented search.



X = O, NOH etc., and a nitrogen hetero-atom at one or more of the positions marked*

The search requires three groups of terms. Group I searches for structures in which the pyrone ring is cited first in the notation and there is at least one nitrogen atom in the other ring. Group II caters for structures in which the pyrone ring contains a nitrogen atom and is cited first. Group III deals with the case in which the pyrone ring is cited first and there is at least one nitrogen atom in the other ring.

Group I:
T66 BO EV
or T66 BO EY
or T66 BV EO
or T66 BY EO
followed by _GN *or* _GK *or* _HN *or* _HK

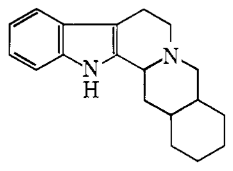
Group II:
T66
followed by _BVN EO *or* _BYN EO
or _BVK EO *or* _BYK EO
or _BON EV *or* _BOK EV
or _BON EY *or* _BOK EY

(and 8 other WLN Fragments)

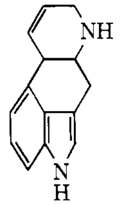
Group III:
T66 BN
or T66 BK
followed by _GO JV *or* _GO JY
(and 6 other WLN fragments)

Some of the cases allowed for are very unlikely. Unfortunately, the ring-frequency list is of little use in a case such as this where the required compounds are very rare and, in all probability, none or very few of the possible notations will be found in the frequency list. A search therefore has to be made for all possible notations, unless chemical knowledge can be used to eliminate some of the possibilities. For example, three or more quaternary nitrogen hetero-atoms are a most unlikely feature of relatively small ring systems.

Searches for Ring Systems Which May Be Fused to Other Rings. A requirement for a ring fragment embedded in a large ring system is often difficult to satisfy as the relevant substructure will no longer have a characteristic stretch of notation. For example the ring systems I and II both contain the carbon-nitrogen skeleton of indole but this is not explicit in the notation of either structure:



I Yohimbine nucleus, WLN is:
T5 F6 D5 C666 EM ON&&TTTJ



II Ergoline nucleus, WLN is:
T C6656 1A P GM LM DUTT&&J

For searches of this type, it is of particular importance that the user should define his interests as precisely as possible. If a user expresses an interest in an embedded ring fragment, irrespective of the character of the complete ring system, the search strategist must consider the possibility that the user's *stated* requirement is expressed in too general terms and is much broader than his *real* requirement. The latter may be restricted to a few specified ring systems for which a relatively simple search strategy can be devised. However, one must also be aware of the danger that this emphasis on a very precise statement of the user's requirements could result in the user being too specific and excluding some *relevant* structures from his stated requirement.

Given that we have defined a user's real requirements, there are two possible approaches to an 'embedded ring' search:

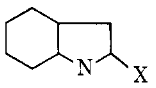
> The user is asked whether he can specify the rings which are likely to be fused to the cyclic fragment of interest. From his knowledge of the subject and its literature he may well be able to identify the fused-ring structures which are more likely to occur. An unfragmented search can then be made for the notations corresponding to each of these structures. The profile will probably perform satisfactorily and give reasonable precision and recall, but it will not retrieve any fused-ring system which the user was unable to predict in advance. Such unpredicted structures may well be of major interest to the user. Indeed, they may be the only ones that are of real interest, since he is already familiar with the chemistry of known structures. This approach was adopted for a user who was interested in all pyrazines, piperazines, and diketopiperazines including bridged structures and complicated fused structures, and there was no evidence that any relevant fused-ring structures had been missed.
>
> A very broad profile strategy is used, which searches for those features which will be common to all the possible notations. For example, a search for ring systems containing the indole nucleus can be made by looking for all compounds containing a five-membered ring, a six-membered ring, and

a nitrogen atom. This can be achieved by using a series of terms such as:

T(n)5(n)6(n)N(n)J
or T(n)6(n)5(n)M(n)J
etc. (using *followed by* logic)

This example illustrates the disadvantage of this type of approach: the features sought are so common that the profile retrieves a very large output of which only a fraction mentions an indole structure. This strategy can be reasonably effective if relevant compounds are characterized by a relatively uncommon structural feature which is always represented by the same fragment or fragments of notation. For example: a ring system containing four sulfur atoms; or a ring system containing a seven-membered ring.

Ring Substituents. WLN is well suited to searches for ring substituents on isolated ring systems. A search for a substituent can be combined with either a fragmented or an unfragmented search strategy for the parent ring system. An example is a search for:



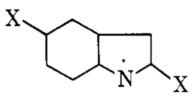
where X is a halogen atom, and both the degree of unsaturation of the ring system and the occurrence of ring carbonyl groups or of other substituents are undefined. In WLN, the position of a substituent is defined by a locant. In all the 2-chloro substituted compounds which would satisfy the above requirement, the chloro substituent will be represented by the notation fragment _CG which will always follow the notation for the ring system. This notation specifies that there is a chlorine atom (G) at position C.

An unfragmented search for the ring system can be combined with a search for the 2-halo substituents as follows:

T56 BNJ
or T56 BM DV CHJ
etc.
followed by _CG *or* _CF *or* _CL *or* _CE

If two substituents are to be specified, then the second list of terms must be modified. Thus for 2,5-halo-substituted compounds, the terms required would be:

_CG(n)GG
_CG(n)GE
_CG(n)GF
_CG(n)GI
_CE(n)GG
etc.



A reasonable value of n would be 10–15, if the emphasis is on recall rather than precision.

Structures containing any substituent at a particular ring position can be retrieved by a search for a particular locant. It is also possible to search for a particular substituent (say chlorine) at any position by specifying:

(space) (any alphabetic character) G

In the case of the serial search programs this can be achieved by the term _(2)G which will search for a space followed within two characters by G. This would also match with G specifying a locant, as well as with G specifying a substituent. An ignore term _G is therefore necessary to prevent the locant match, and an additional term _GG to cover the one instance where the ignore term would prevent a useful match. The circuitous character of this logic illustrates the need for a simple means of distinguishing between locant and nonlocant alphabetic characters (this facility is incorporated in the bit screens used by the RADIICAL programs[9]).

Searches for Acyclic Substructures. To a first approximation, the ease with which an acyclic substructure requirement can be satisfied is determined by the number of branch points with undefined valence bonds in the fragment of interest. If none occur, there will be only a few, well-defined notation paths through the fragment, and the search is straightforward. If the fragment contains one or more branch points, however, the search poses many problems. Not only are there a large number of possible notation paths through the fragment, but some atom-to-atom connections may not be readily apparent to a machine search. In the cases where two or more branch points occur, a satisfactory search can only be made if the possible notations representing the fragment have very characteristic features.

An example of a requirement for an acyclic substructure with no branch points is:

X—NH—CO—CH=NH—X
where X = any group other than H

This fragment will be coded as MV1U1M or M1U1VM depending on the nature of the undefined groups. There are thus only two possible notation paths, and a search for this fragment would be:

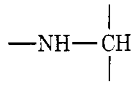
MV1U1M
or M1U1VM

Searches for many monovalent or divalent groups require only one or two search terms:

Group	Search Term
—OH	Q
$—NH_2$	Z
—COOH	VQ or QV
$—CF_3$	FXFF or XFFF
—C(=O)—	V
—NH—	M
—C(=O)—O—	VO or OV

WLN is obviously well suited to searches for such functional groups. Furthermore, a manual search for these groups in a conventional index of systematic nomenclature will often give unsatisfactory results. In practice, most searches for such simple functional groups are restricted by another structural requirement. For example, a user may be interested in hydroxyl derivatives of phenanthrene.

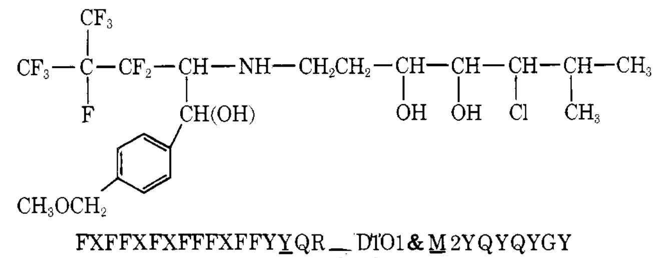
An example of a requirement for a substructure with one branch point is:



There are six possible notation paths through this fragment, and the possible notations are MY, YM, or Y followed within an undefined number of characters by M. The required search strategy is:

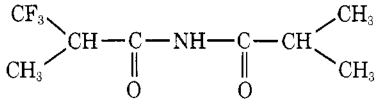
MY
or Y(n)M

This term can never achieve perfect recall and precision. A low precision will be obtained if n is large (say 10), and the recall will be low if n is small (say 3). It is usually necessary to accept a compromise in which an intermediate value of n is chosen. The following example shows how both recall and precision failures occur with this type of search, when an intermediate value of n = 7 is chosen:



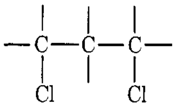
FXFFXFXFFFXFFYYQR _ D1O1& M2YQYQYGY
1 2 3 · 4 5 6 7 8

This compound contains the required fragment but no match will be made because eight characters intervene. On the other hand, though the following compound does not contain the required substructure, a match will be made with the notation as only two characters intervene between the Y and the M.

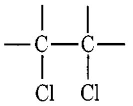


FXFF$\underline{Y}\underset{1}{\&}\underset{2}{V}\underline{M}$VY

As the number of branch points with undefined valence bonds increases, the difficulties of devising satisfactory search strategies become more serious. An example is a requirement for the following substructure:

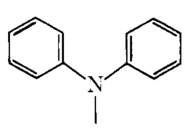


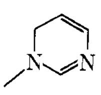
The number of notation paths through this fragment is very large, and the problem of devising a search which will match with the notation representing this fragment and will not match with the notations representing structures such as

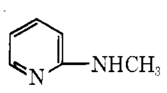


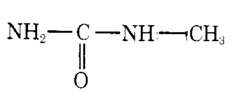
is insoluble. In such cases, the user should be asked whether he can be more specific in stating his requirements, for instance, by suggesting the groups which are likely to be attached to the unspecified bonds. In this approach to the problem there is an obvious danger that the user's requirements will be distorted to obtain a reasonable search. Even if the user can specify some of the groups which are attached at the branching points, this may not always be of assistance in devising a search strategy. It is, for example, of little use to know that the unknown groups will be alkyl groups, as these groups themselves contain an unknown degree of branching. However, a reasonably effective search may be possible if the relevant structure fragment possesses relatively uncommon features. For example, although the fragment —N—C—C≡C— contains two branch points with undefined valencies, we can devise a strategy which does not produce an excessive output because the acetylenic bond occurs in only a few registered compounds.

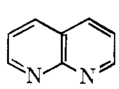
Searches for Substructure Fragments Which May Form Part of Both a Cyclic and an Acyclic Part of the Molecule. Only five examples were encountered in the present project of a requirement for an array of atoms, irrespective of their molecular environment. In future, there may be a greater demand for searches of this type as understanding of structure-function relationships improves. An example would be a search for the fragment —N—C—N—C— which would be satisfied by any of the following compounds.











There is no adequate method of satisfying such requirements by means of WLN character matching, because the encoding rules make a primary distinction between ring and chain structures. Only a connection-table (or possibly a fragment-code) search could retrieve substructure fragments of this type. The only possible approach in a WLN string search is to attempt to determine all likely substructures containing the required fragment. This approach would almost always give poor results, since no equivalent of the ring-frequency list is available as a guide in such cases.

It is possible that future users, who are interested in structure-function relationships, will wish to define their requirements more precisely than in the example given above (—N—C—N—C—) by specifying the *electronic* or *stereochemical* environment of relevant fragments (noted by C. E. Granito, ISI). Such characteristics are not expressible in WLN, as it now exists, and could only be incorporated into a search strategy if they can be used to define specific atomic arrays which are relevant to the users.

SUBSTRUCTURE SEARCH STRATEGIES—CONCLUSION

Searches of a structure file coded in WLN are:

(a) Very effective for isolated ring systems and for acyclic structures in which there are no branch points with undefined valence bonds

(b) Moderately effective for embedded ring systems and for acyclic structures with one branch point with undefined valence bonds, but a substantial amount of intellectual effort is required in profile construction

(c) Only effective for acyclic structures with more than one branch point with undefined valence bonds, if the fragment has very characteristic features

(d) Almost never effective for fragments for which no molecular environment can be defined

The difficulties encountered in searching for isolated rings within a fused ring system and for acyclic fragments having one or more branching points do not result from inadequacies in the search logic but are due to inherent limitations of a character-matching search of Wiswesser notations. These difficulties could only be avoided by the use of a different type of structure file in which all atom-to-atom connections are explicit in the computer record. Thus, the Crossbow system used by ICI Pharmaceuticals[3] converts the WLN records used for input into a connection-table record to allow for effective searches for all types of substructure requirements. On the other hand, a notation search may be simpler, quicker, and cheaper than other, more sophisticated, methods.

The potential value of a WLN character-matching

Table III. Classification of Structural Requirements in Relation to Type of WLN Search

Cyclic	
Isolated ring systems	42
Embedded ring systems	
(i) all possible systems specified	17
(ii) possible ring systems not specified	32
Acyclic	
No branch points with undefined valence bonds	10
One such branch point	10
More than one such branch point	18
Fragments which may occur in both cyclic and acyclic environments	5

Categories are not exclusive; some users had two (or more) distinct interests which were included in two or more categories.

search as a means of substructure retrieval will depend on the type of requirements submitted to a search system. For example, if a majority of "real" requirements submitted by a representative user population are for fragments for which no molecular environment can be defined ((d) above), WLN will be an ineffective tool for substructure retrieval. Table III shows a classification of the subject requirements of 101 users according to the type of search strategy required. We were able to develop reasonably effective strategies for the majority of these users.

ACKNOWLEDGMENT

We wish to express our thanks for the help given by A. K. Kent and I. C. McCracken of the United Kingdom Chemical Information Service; by the Institute for Scientific Information and, in particular, A. E. Cawkell, C. E. Granito, and M. Rosenberg; and by the Oxford University Computing Laboratory. We also wish to thank the research workers who comprised our user population. The project was funded by a grant from the Office for Scientific and Technical Information (U.K.). The grant-holder was Sir Ewart Jones, F.R.S., Oxford University.

LITERATURE CITED

(1) National Academy of Sciences, Publ. No. 1733, "Chemical Structure Information Handling: A Review of the Literature: 1962-1968," Washington, D. C., 1968.

(2) Lynch, M. F., Harrison, J. M., Town, W. G., and Ash, J. E., "Computer Handling of Chemical Structure Information," Macdonald and Co. (Publishers) Ltd., London, and American Elsevier Publishing Company Inc., New York, 1971.

(3) Campey, L. H., Hyde, E., and Jackson, A. R. H., "Interconversion of Chemical Structure Systems," *Chem. Brit.* **6,** 427-30 (1970).

(4) Dammers, H. F., and Polton, D. J., "Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures," *J. Chem. Doc.* **8,** 150-60 (1968).

(5) Bowman, C. M., Landee, F. A., Lee, N. W., Reslock, M. H., and Smith, B. P., "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File," *Ibid.*, **10,** 50-54 (1970); and earlier work by these authors cited therein.

(6) Leiter, D. P., Morgan, H. L., and Stobaugh, R. E., "Installation of a Registry System for Chemical Compounds," *Ibid.*, **5,** 238-42 (1965).

(7) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Calderon, M. M., and Warner, A., "*The Index Chemicus Registry System:* Pragmatic Approach to Substructure Chemical Retrieval," *Ibid.*, **10,** 54-8 (1970).

(8) Rössler, S., and Kolb, A., "The GREMAS System, an Integral Part of the IDC System for Chemical Documentation," *Ibid.*, **10,** 128-34 (1970).

(9) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.

(10) Palmer, G., "Wiswesser Line-Formula Notation," *Chem. Brit.* **6,** 422-6 (1970).

(11) Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "The Evaluation of a Current-Awareness Service based on the *Index Chemicus Registry System,*" in preparation.

(12) Crowe, J. E., Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "Development and Evaluation of a Current-Awareness Service based on the *Index Chemicus Registry System,*" The Experimental Information Unit, Oxford University. Report to the Office for Scientific and Technical Information (U.K.), June, 1973.

(13) 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,** 106-10 (1971).

The *Integrated Subject File.* I. Data Base Characteristics

W. C. ZIPPERER, R. E. STEARNS, Jr., and M. K. PARK*
Computer Center, University of Georgia, Athens, Georgia 30602

Received October 11, 1972

Characteristics of Volume 71 of the *Integrated Subject File* (ISF), the computer-readable data base corresponding to the *Chemical Abstracts Subject* and *Formula Indexes,* are reported. Minimum, maximum, and average lengths and frequency counts for the data elements in the Chemical Substance and General Subject segments of the Standard Distribution Format (SDF) files distributed by Chemical Abstracts Service are presented. Similar data are tabulated for the same files as converted for use with the UGA Text Search System and for a merged data base created from the index entries from the ISF and the bibliographic information from *CA-Condensates.*

As a part of its continuing evaluation of computer-readable bibliographic data bases, the University of Georgia (UGA) has begun a research study on the *Integrated Subject File* (ISF), the computer-readable data base generated by Chemical Abstracts Service (CAS) which corresponds to the semiannual *CA Subject and Formula Indexes.* This study is designed to evaluate alternative file organizations, data element content, and search strategies for the ISF data base, as well as combinations and comparisons with *CA-Condensates,* the corresponding bibliographic data base.[1] The project description defines seven major tasks for investigation over an 18-month period. These tasks include: creation of bibliographic data base search files, collection and characterization of questions, comparison of biblio-

* To whom correspondence should be addressed.