(2) Fritts, L. E.; Pollack, N. M. "Chemical Biological Inquiry System—An Online System for Searching Chemical Structures in an Interactive Mode", National Meeting of the American Chemical Society, Honolulu, Hawaii, April 1979; American Chemical Society: Washington, DC, 1979.
(3) Pollack, N. M.; Fritts, L. E. "Usage of the Diamond Shamrock Chemical-Biological Inquiry System", National Meeting of the American Chemical Society, Honolulu, Hawaii, April 1979; American Chemical Society: Washington, DC, 1979.
(4) Sorter, P T.; Granito, C. E.; Gilver, J. C.; Gelberg, A.; Metcalf, E. A. "Rapid Structure Searches *via* Permuted Chemical Line Notations". *J. Chem. Doc.* **1964**, *4* (1), 56-60.
(5) Chemical Information Management Inc., P. O. Box 2740, Cherry Hill, NJ 08034.
(6) ROSCOE is a comprehensive software system for online program development with multiple interactive support functions. ROSCOE is the property of Applied Data Research, Inc., Princeton, NJ.
(7) Ash, Janet E.; Hyde, Ernest. "Chemical Information Systems"; Wiley: London, 1975; pp 227-242.
(8) Mark IV Systems is an applications development system by Informatics Inc., 21050 Vanowen Street, Canoga Park, CA 91304.
(9) Fraser Williams (Scientific Systems), Ltd., Glendower House, London Road South, Poynton, Cheshire SK12 1NJ, England.

Wiswesser Line Notation as a Structural Summary Medium†

TRISHA M. JOHNS*
G. D. Searle & Co., Skokie, Illinois 60076

MICHAEL CLARE
G. D. Searle & Co., Ltd., High Wycombe, Bucks, England

Received August 24, 1981

Wiswesser Line Notation (WLN) is well established as a technically unambiguous and efficient method of denoting chemical compounds. Its true appeal, however, lies in the fact that it is a linguistic rather than a merely symbolic notation. The syntactical WLN can be broken down into easily recognizable wordlike fragments retaining much of the original information content and yet often more immediately meaningful than the full molecular description, especially for the purpose of identifying common structural features. This paper describes how G. D. Searle is using such WLN fragments as an integral part of a minicomputer-based WLN-oriented data base system designed to handle its internal compounds. Overall system features are discussed, including report generation, data entry, and search capabilities.

INTRODUCTION

Decentralization of computer systems at G. D. Searle led to a need by the Information Services Department to redesign the information system that handles internal chemical structures. A batch retrieval system had been developed for the Honeywell 6060; now a more flexible network of DEC minicomputers was available. Through years of usage, we had come to understand not only the idiosyncrasies of the file but also the capabilities we required. The historical base we had to work with was an existing computer-readable file of Searle compounds expressed in WLN, which was searchable in the batch mode by using a combined bit-screen, character-by-character search or manually from alphabetic listings and updated with keypunch cards. This paper will describe the new online system we are developing, based on WLN and run on VAX/780 and PDP 11/40 minicomputers.

WISWESSER LINE NOTATION

That our files are in WLN proved to be more of an asset than a constraint when we considered the design of an interactive system. The linguistic nature of the WLN readily lends itself to a more human-engineered system, one that can be searched efficiently at varying levels of specificity. The WLN is an unambiguous representation of the compound, made up of wordlike groups of symbols arranged by definite rules of syntax. Like words from a sentence, substructural fragments can be described out of the context of the full WLN string. The ranking of symbols and application of WLN encoding rules result in unique notations but at the same time can serve to obscure the direct comparison of substructural features in different compounds, mainly because of "head-to-tail" inversions. An example of this problem is shown in Figure 1. The WLN representation of urea derivatives is legitimate in any of these symbol arrangements. Should either of the nitrogens be part of a ring system, or be substituted, the urea designation is lost within the WLN string.

Direct substructure searches of such WLN data bases must take into account the head-to-tail inversion problem, either by utilizing conditional logic to search for these alternative representations when a character search is employed or by the generation of a connectivity table followed by a partial atom-by-atom comparison. In either case, the WLN records to be analyzed are usually preselected by some bit-screen classification, to avoid having to process all the records.

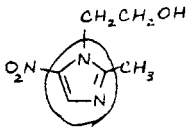
We have adopted a different approach, based on an adaptation of a WLN fragmentation method suggested by C.E. Granito of CIMI.[1] This fragmentation scheme focuses attention on ring systems and common functional groups, ignoring simple aliphatic chain linkages. It is important to note that the fragments are derived from the chemical structure itself, not its WLN. This substructural description is written by using WLN symbols, ordered by the latest end alphabetically. Figure 2 shows how the fragments are related to the structure drawing. The compound, metronidazole, can be viewed as an imidazole ring with various substituents. In addition to the ring system itself, the hydroxy, nitro group, and ring nitrogens are the fragments for this compound, since the aliphatic carbons are ignored. In the language of WLN, the fragments are T5N CNJ denoting imidazole and Q denoting a hydroxy group. The "space hyphen T" indicates that

† Presented at the Second Chemical Congress of the North American Continent, Las Vegas, NV, Aug 26, 1980.

0095-2338/82/1622-0109$01.25/0

MVM	ZVM	NVM	N1&VN etc.
MVN	ZVN	NVN	N2&VN etc.
MVZ	ZVZ	NVZ	NR&VN etc.

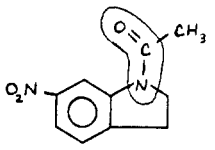
Figure 1. Partial list of legitimate WLNs for the urea fragment.



WLN: T5N CNJ A2Q B1 ENW

FRAGMENTS: T5N CNJ
Q
NW -T
N
N -02

Figure 2. WLN fragment example.



WLN: T56 BNT&J BV1 HNW

FRAGMENTS: T56 BNT&J
NW -T
VN

Figure 3. Relation of fragments to structure.

the nitro group (NW) is attached to a heterocyclic ring. Each nitrogen atom in the ring is a fragment; "space hyphen 02" means that two of them are present in the molecule.

Figure 3 points out how the fragments are derived from the structure and not from the syntactical WLN, thus avoiding the head-to-tail inversion problem. The fragments are open ended in the sense that any substructural unit can be denoted in WLN, and so the particular set of fragments in the file can be expanded when necessary. For instance, we originally had excluded phenyl rings but realized that in our files the phenyl group is important; so this was subsequently added as a legitimate fragment.

The WLN fragments form the basis of our substructure search system. While utilizing conventional WLN symbols, the fragments are meaningful on sight to anyone versed in the WLN language, and they are written such that a given substructural unit corresponds to a unique WLN representation by the consistent ordering of symbols in the same direction.

Each of the fragments thus produced is used as an index pointer to the record from which it was derived, so in a sense one is producing a "keyword index", though the original "text" need not contain each word explicitly.

TRAINING THE END USER

The one drawback of an information system based on a secondary language such as WLN is that the end user, i.e., the bench scientist, usually does not know the language. We have found that the WLN fragments can serve as a useful educational aid to familiarize new users with the basic elements

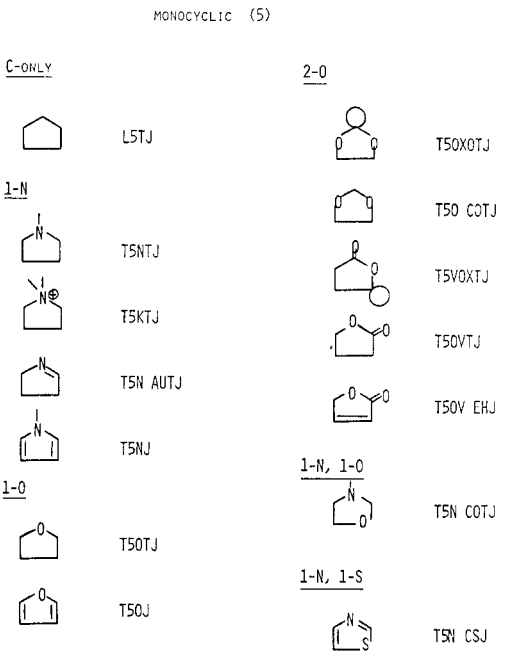


Figure 4. Page from user manual.

of the WLN. Figure 4 shows a sample page from our fragment manual. Our data base of 50 000 WLNs produced approximately 250 000 fragments, and of these, about 5000 are unique and are represented in the user manual. The user can peruse the manual, selecting those ring systems and functional groups of interest. Through the fragment search program, the user can scan the data base for structural types or do a specific substructure search. The search program was designed with the end user in mind, demanding a minimum amount of computer familiarity to achieve good results. For example, the search strategy itself is optimized by the computer for the fastest retrieval. As there is no online structure display at present, the user is referred to a microfiche file for two-dimensional structure drawings. Users have the capability of scanning, editing, and listing the hit files to produce amended versions or polished hard copies of search results. The hit files can also be passed on to other user programs, providing a means of linking the structure file to other Searle data bases.

FRAGMENT SEARCH PROGRAM

The fragment search program is designed to run in a real-time interactive environment and is currently available on a number of machine types in our network, including the VAX/780 and PDP 11/40. The present version is written in BASIC-PLUS 2, which is available on both of the series. Future plans include upgraded versions to be written in FORTRAN. The search program operates in five phases: (1) setting up the compound threshold, (2) entering the search profile, (3) optimization by the computer of the search strategy, (4) the search itself, and (5) displaying the search results.

An example of a search for prostaglandins is shown in Figure 5. After logging on to the system, the fragment search program is invoked, and the question is asked if search limits should be set. All user responses are underlined to highlight how little input is required by the user. "Search limits" refers to the various classifications of internal compound numbers within Searle, and this first option allows the user to restrict subsequent operations to a fixed window of the whole set. For this example, the search will be limited to those compounds used as standards in biological testing that are available from outside sources, the E range.

```
RUN GENKEY
ANSWER "YES" TO SET SEARCH LIMITS? YES
CURRENTLY THE DATABASE COVERS THE FOLLOWING RANGES
00001 - 99999
B0001 - B9999
C0001 - C9999
E0001 - E9999
R0001 - R9999
W0001 - W9999
X0001 - X9999
ENTER LOWER THRESHOLD? E0001
ENTER UPPER THRESHOLD? E9999
ENTER FRAGMENT 1? L5TJ
OCCURS 420 TIMES IN THE DATABASE
ENTER FRAGMENT 2? VQ:VO
     5,884 OCCURENCES OF VO
     3,201 OCCURENCES OF VQ
OCCURS 9,085 TIMES IN THE DATABASE
ENTER FRAGMENT 3?
ANSWER "YES" TO PROCEED WITH SEARCH? YES
SEARCH TIME OF    8 SECONDS-   8 HITS
ANSWER "YES" TO CONTINUE? YES
ENTER "F" TO PRINT FULL HIT LIST? F
ENTER A FILENAME FOR THE HITS FILE? PGF

    1)    E0218
L5TJ AQ B6VQ C1U1YQ5 DQ -A&ABD -B&C -A -T
    2)    E0443
L5TJ AQ B6VQ C1U1YQ5 DQ -A&BD -B&AC -A -T
    3)    E0500
L5TJ AQ B2U4VQ C1U1YQ5 DQ -A&ABD -B&C -A -CT
    4)    E0503
L5TJ AQ B2U4VQ C1U1YQ5 DQ -A&BD -B&AC -A -CT
    5)    E0764
L5TJ AQ B2U4VQ C1U1YQ2U3 DQ -A&ABD -B&C -A -CTC
    6)    E0766
L5TJ AQ B2U4VQ C1U1YQ2U3 DQ -A&BD -B&AC -A -CTC
    7)    E0767
L5TJ AQ B2U4VQ C1U1YQOR CG& DQ -A&ABD -B&C -S -CT
    8)    E0835
L5TJ AQ B2U4VQ C1U1YQ1OR CXFFF& DQ -A&ABD -B&C -S -CT

8 DISTINCT HITS IN PGF - ANSWER "YES" TO SAVE IT? YES
```

Figure 5. Sample search 1.

The first fragment, L5TJ, representing a five-membered fully saturated carbocyclic ring, is entered. The program responds that this fragment occurs 420 times in the complete data base. As the search profile is built, the number of occurrences can be a good indicator of needed search refinement. A simple "or" condition is entered next (indicated by a colon) indicating that an acid (VQ) or ester (VO) function is required in the molecule. The program will continue to ask for additional fragments until a carriage return is hit. At this point, the search can be aborted by answering "no" to the continue question.

No matter what order the search profile is entered, the program performs the actual search in the most efficient route as determined by the relative frequency of the fragments in the data base. This information is maintained along with the unique fragment types in an indexed fragment directory.

The response after the search phase is completed gives the actual search time and number of hits. If the resulting number of hits is too high and refinement of the search is necessary, the answer need not be displayed. The response to the next question determines whether the full WLN and compound number is to be printed or whether a list of compound numbers is sufficient. These hits are automatically written to a hit file, which the searcher is then asked to name. The name can be an existing file name or a new one. For example, one may wish to do several separate searches and combine the results. Duplicate hits are automatically eliminated in the combined hit file. In this case, the WLNs for eight prostaglandins and their internal accession numbers are displayed. Depending on the response to the final question, the hit file will either remain in the user's account on exiting from the program or will be deleted.

```
                HITS FILE OPTIONS

BROWSE:     DISPLAYS COMPOUND NUMBER AND WLN OF EACH
            RECORD IN THE HITS FILE, ARRANGED IN COMPOUND
            NUMBER ORDER.  "CARRIAGE RETURN" DISPLAYS THE
            NEXT RECORD.  "D" DELETES THAT RECORD.


PRINT:      SENDS EDITED HITS FILE TO LINE-PRINTER FOR
            HARD-COPY SEARCH RESULTS.

KILL:       ERASES THE HITS FILE.

EXIT:       EXITS FROM THE PROGRAM WITHOUT AFFECTING THE
            HITS FILE.
```

Figure 6. Hit file options.

```
ENTER FRAGMENT 1? NVN:NVM
    372 OCCURENCES OF NVM
    202 OCCURENCES OF NVN
OCCURS 574 TIMES IN THE DATABASE
ENTER FRAGMENT 2? T+
ALL SUCH ENTRIES INCLUDED
ENTER FRAGMENT 3? #R
OCCURS 20,322 TIMES IN THE DATABASE
ENTER FRAGMENT 4?
ANSWER "YES" TO PROCEED WITH SEARCH? YES
SEARCH TIME OF   4 SECONDS  -  2 HITS
ANSWER "YES" TO CONTINUE? YES
ENTER "F" TO PRINT FULL HIT LIST? F
ENTER A FILENAME FOR THE HITS FILE? UREAS

    1)  E0465
T6MVNJ DZ EF

    2)  E0334
T6N DNTJ AVN2&2 D1

2 DISTINCT HITS IN UREAS -  ANSWER "YES" TO SAVE IT? YES
```

Figure 7. Sample search 2.

HIT FILE

The hit file may be edited by using program options shown in Figure 6. Each compound number and associated WLN can be displayed one by one, allowing the user to delete or save them individually. In this sample search, for example, a five-membered carbocycle is likely to be a prostaglandin in our data base, but this parameter would not be sufficient when searching a file of more general organic compounds. Had the search come up with irrelevant hits, the hit file would be scanned online and only the good ones saved. The print command sends the output to a line printer for hard-copy search results, and the hit file can then be saved or deleted.

The hit file may be searched in more depth, i.e., character by character, through a purchased DEC software package called Datatrieve.[2]

OTHER SEARCH OPTIONS

The fragment search may be done on partial fragments with or without Boolean modifiers as shown in sample search 2 in Figure 7. The first expression asks for a urea derivative, with either one or both nitrogens disubstituted. The right hand truncation capability is exemplified by the root T+, indicating any heterocyclic system. When truncation results in a large

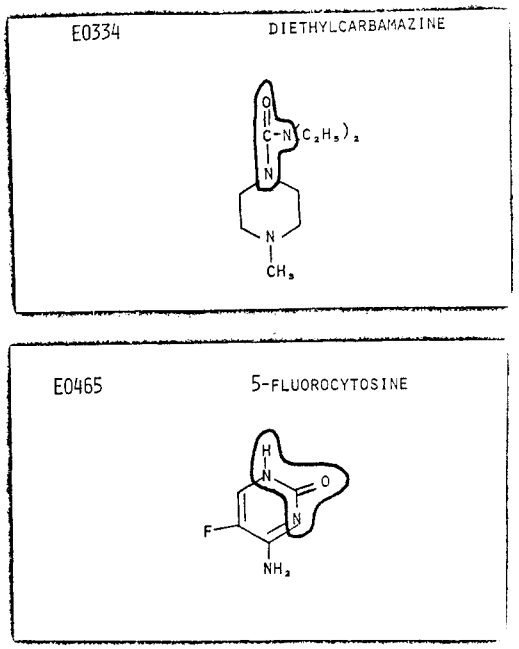


Figure 8. Structure cards.

```
RUN WLNIN
ENTER 6-CHARACTER FILENAME PREFIX: UPDATE
NEW FILE UPDATE.WLN BEING CREATED
ENTER WLN: T C566 DO FO JN MV EHJ J2 L1
WLN ENTERED: T C566 DO FO JN MV EHJ J2 L1
IS DATA CORRECT? YES
WLN NOT IN FILE, ENTER COMPOUND TYPE: E
      E0957  T C566 DO FO JN MV EHJ J2 L1
WLN APPENDED TO UPDATE.WLN
ENTER WLN:
```

Figure 9. WLN entry.

number of possibilities, as in this case, the number of possibilities and their occurrences is suppressed. An example of "NOT" logic is shown by the third fragment, #R, indicating the exclusion of phenyl rings.

Another carriage return and a "yes" to proceed results in two hits in a search time of 4 s. Figure 8 shows the structure cards that the user would be referred to for a structure drawing. A urea fragment appears in both compounds, although in one case the fragment is wholly contained within the ring, and in the other case, only one element is in the ring.

The hit file may be linked to other Searle data bases, through DEC's Datatrieve software. For example, a search of the biology data associated with the hit file from a substructure search is done by a Datatrieve procedure. The search can be limited to a number of parameters (such as type of test or date ranges). New Datatrieve procedures are easily written to change parameters or report formats and also to link other data bases with common access points.

WLN/WLN FRAGMENT FILES

All of the data structures in the combined WLN/WLN fragment system are maintained as indexed sequential files accessed under control of the DEC RMS file handling system, and their total storage requirement is typically 20 Mbytes of a 300-Mbyte disk.

The full WLN file has a fixed record length of 155 bytes. The primary key of the internal structure file is the compound

```
RUN FRAGIN
ENTER 6-CHARACTER FILENAME PREFIX: UPDATE
NEW FILE UPDATE.FRA BEING CREATED
ENTER COMPOUND NUMBER: 10295
RING FRAGMENT - T5N CNJ
ENTER FRAGMENT: Q
ENTER FRAGMENT: NW +T
ENTER FRAGMENT: N
ENTER FRAGMENT: N -02
ENTER FRAGMENT:
FRAGMENTS ENTERED FOR 10295

      1) T5N CNJ
      2) Q
      3) NW +T
      4) N
      5) N -02

IS DATA CORRECT? NO
ENTER LINE NUMBER AND CORRECTION
3,NW -T
FRAGMENTS ENTERED FOR 10295

      1) T5N CNJ
      2) Q
      3) NW -T
      4) N
      5) N -02

IS DATA CORRECT? YES
FRAGMENTS APPENDED TO UPDATE.FRA

ENTER COMPOUND NUMBER:
```

Figure 10. WLN fragment entry.

```
RUN FORMIN
ENTER 6-CHARACTER FILENAME PREFIX: UPDATE
TYPE CR TO CONTINUE:
ENTER COMPOUND NUMBER: 18052
ENTER MAIN MOLECULAR FORMULA: C25H39N1O2
ENTER FORMULA FOR ADDUCT #1: C2H2O4
ENTER MULTIPLICATION FACTOR: 1
ENTER FORMULA FOR ADDUCT #2: H2O
ENTER MULTIPLICATION FACTOR: 0.5
ENTER FORMULA FOR ADDUCT #3:
      MW = 484.63   C25H39NO2  :  C2H2O4  :  0.5 H2O
ENTER COMPOUND NUMBER:
```

Figure 11. Molecular formula entry.

number. Included in each record is the molecular formula, molecular weight, source code, bit screen profile, and full WLN.

DATA ENTRY PROGRAMS

All data entry and file maintenance is done online. As the volume of new entries is not large at any given time, all WLNs, fragments, and molecular formulas are entered as the structures are submitted. Figure 9 shows a typical WLN entry session. The entry is repeated by the computer for verification. A "NO" response indicating that the data is not correct will have the "ENTER WLN" prompt repeated. If the WLN is in the file already, its accession number is printed; otherwise the message "enter compound type" is given. The classification is entered, the next highest number is assigned automatically, and the record is added to the file. The data entry programs are interim procedures as we are investigating use of a screen-fill software package with built-in validity checks for the future.

Figure 10 shows the WLN fragment entry scheme. When the compound number is entered, the program automatically pulls out the ring fragment from the full WLN, as this is typically the longest entry and remains unchanged as a fragment. The rest of the fragments are determined by the information scientist, who enters one fragment after each

prompt. A carriage return indicates that no more fragments are forthcoming, and all the fragments are immediately displayed for verification. When an error is made, a simple entry indicating the line number followed by the correct fragment corrects the entry. The fragments are then displayed again for verification, and a "YES" response signals the computer to add this record to the fragment file.

There are few rules for molecular formula entry as shown in Figure 11. The formula can be entered in any order. Single atoms must be followed by a one. Formulas for adducts are added separately, with a multiplication factor. The total molecular weight is calculated automatically and printed out with the entry in the standard Hill form, followed by the adducts as shown. Eyeball verification of the formula is corroborated by a comparison of the weight to that calculated by the submitting chemist.

FUTURE PLANS

The system has been built in a modular fashion, beginning with the development of the fragment file and search programs. These modules have been operational over a year, and we have found that the vast majority of our substructure searches can be done at this level. The next important modules to be addressed are atom-by-atom searching and online structure display.

The thrust of the decentralization effort has been to put information into the hands of the end users. The system described is an attempt to give them the opportunity to search the internal structure data base with a minimum knowledge of WLN and computer jargon until such a time that direct structure searching can be engineered. In addition to the fragment manual, we have begun a series of tutorials for our research staff, including hands-on problem sessions. An international network will bring these programs into the laboratories of Searle research scientists throughout the world.

REFERENCES AND NOTES

(1) C. E. Granito, CIMI, P.O. Box 2740, 411 Route 70 East, Cherry Hill, NJ 08034.
(2) "Datatrieve-11 V2.0. User's Guide"; Digital Equipment Corporation: Maynard, MA, 1980.