

Using the Wiswesser Line Notation (WLN) for Online, Interactive Searching of Chemical Structures[†]

LOIS E. FRITTS* and MARGARET MARY SCHWIND

Diamond Shamrock Corporation, T. R. Evans Research Center, Painesville, Ohio 44077

Received August 24, 1981

An online, interactive chemical information system is described in which chemical structures are retrieved by WLN or functional fragments derived from the WLN. Examples are given of structure and substructure searches in which functional fragments are combined to retrieve analogous compounds. Procedures are described for integrating the results of chemical searches with biological data to contribute to research in the biosciences.

Wiswesser Line Notation (WLN)¹ is a line formula notation which enables a three-dimensional structure to be stored for computer searching and manipulation. The Diamond Shamrock Online Inquiry System,^{2,3} an interactive chemical information system, uses WLN as the primary tool for registry of chemicals and for the storage and retrieval of chemical and physical data.

Information chemists at Diamond Shamrock were among the first to use the WLN for encoding chemical structures into machine-readable records for the retrieval of chemical structures and substructures.⁴ For many years, the WLN records were used to prepare printed, permuted indexes in which related ring systems and substructural units appeared together in an alphabetic listing. A data line which summarized the biological activity of each compound alternated with the WLN print line to draw attention to any correlation in biological activity with the substructural unit.

As the number of compounds in the file increased rapidly, the permuted indexes became cumbersome and were often out of date by the time they were issued. In the mid 1970s, plans were formulated to modernize the chemical file into an online interactive search and registry system, utilizing the most recent technological advances, yet retaining the foundation of the WLN structure file. The goal was to build a system which could be efficiently maintained, would correlate chemical and biological information, and would be readily available to the end user.

The chemical portion of the online system was designed and implemented by Chemical Information Management, Inc. (CIMI).⁵ The search capabilities of the system allow the online interactive searching of our data base of nearly 50 000 compounds to be performed accurately and efficiently within a matter of seconds. Figure 1 presents a simplistic overview of the online data base design. The master data file contains all of the chemical and physical information which we have chosen to store for each compound in the system; compound number, source code and number, project number, registration date, WLN, chemical name, empirical formula, melting or boiling range, refractive index, known or suspected hazards, and solubilities. The fragment master file stores WLN fragments which are created by the FRAGEN program, an algorithm which is proprietary to CIMI. This program reads the WLN, generates fragment codes from the notation, and stores and indexes them in the fragment master file. The external hit file can be used to store compound registry numbers for those compounds which have been retrieved from a search. WLN fragments, as defined by CIMI, fall into two categories, ring systems and functional groups. Ring systems are the codes depicted in WLN as within the symbols T—J or L—J. A simple benzene ring (the R symbol) is not con-

sidered to be a ring system by this program. Functional groups to the CIMI program are defined as heteroatoms and selected cases of carbon, such as the C in CN for the linear carbon in a cyano group, V for a carbonyl, and SY for a thiocarbonyl. The fragment is built up of connecting heteroatoms until a carbon is encountered.

Although connecting atoms in a molecule are frequently found as noncontiguous symbols in a WLN string, they will be contiguous in a fragment. For example, the structure in Figure 2, 3-chloro-*N,N*-dimethyl-1-piperidinecarboxamide, has the WLN notation T6NTJ AVN1&1 CG. In this WLN, the N symbol, representing nitrogen within the ring system T6NTJ, is not contiguous to the VN symbols which represent the carbonyl and nitrogen of the carboxamide group. However, the program which generates the codes from the WLN does recognize the connectivity of the atoms and connects the symbols as NVN.

If a functional fragment is attached to a ring, it is noted separately from fragments which are not attached to ring systems. For example, in Figure 2, G is the WLN symbol for chlorine and this symbol with the code -T indicates that chlorine is substituted directly onto a heterocyclic ring. This differentiates it from chlorine atoms attached to alkyl groups or other kinds of ring systems. In Figure 3, the chlorine in chlorobenzene is represented by G -R, the chlorine in chloronaphthalene by G -L, and the chlorine in chloropyridine by G -T. This allows for very specific searching as well as for general searching where a chlorine atom is present in a compound. Essentially all molecules with chlorine attached to a carbon could be retrieved by the search command "G or G -T or G -R or G -L". However, in our modest file, as many as 6000 compounds might be retrieved. By limiting the search to chlorines connected to a heterocyclic ring, about 1000 would be retrieved. If a unique heterocyclic ring, such as pyridine, were specified, only about 200 compounds would be retrieved. Hits could be limited still further by the use of additional fragment codes, using simple Boolean Logic.

One of our goals in designing the online inquiry system was to make the chemical data readily available to the end user. We believe that our system is simple enough that even a novice can gain valuable information from the system. A user of the system must first log on with an identity code and a password which allows him access to certain portions of the system and may restrict him from other administrative functions. After logon, a menu, as in Figure 4, appears which prompts the user to indicate which function is requested. At Diamond Shamrock one might search for materials which are related to chlorothalonil, tetrachloroisophthalonitrile, a commercial broad-spectrum fungicide which was developed at Diamond Shamrock's Corporate Research Center. By entering NC -R, representing one or more cyano groups on a benzene ring, several hundred compounds would be found. The search parameters can be restricted to a specific request for two cyano

[†] Presented at the National Meeting of the American Chemical Society, Las Vegas, NV, Aug 1980.

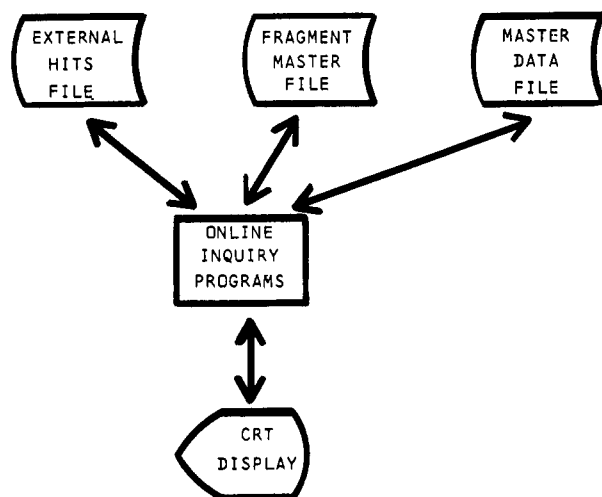


Figure 1.

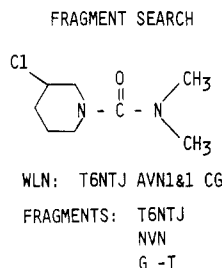


Figure 2.

CHLORINE FRAGMENTS

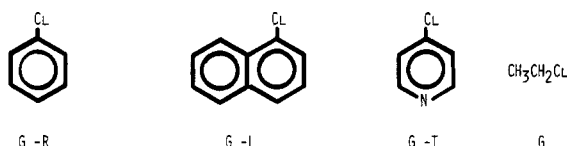


Figure 3.

CHOOSE ONE OF THE FOLLOWING FUNCTIONS BY ENTERING YES:

SEARCH:

LOOK UP COMMAND BY COMPOUND NUMBER, WLN, OR EMPIRICAL FORMULA

FIND THE CPDS WITH SPECIFIC

FRAGMENTS	FROM HITS	OR MASTER
TEST VALUES	FROM HITS	OR MASTER
MULTIPLE CRITERIA	FROM HITS	OR MASTER

UTILITY:

GET AN EXTERNAL HITS FILE FROM AN EARLIER SEARCH

SORT CURRENT HITS FILE

DISPLAY CURRENT HITS FILE

RETRIEVE CATALOG OF EXTERNAL HITS FILE

ERASE HITS FILE

EXHIBIT STATUS OF PENDING UPDATES

HELP

UPDATE (ADMINISTRATIVE OR CLERICAL USERS ONLY)

Figure 4.

groups on benzene rings and four chlorine groups on benzene rings. Now only a few compounds would be retrieved, which can be displayed by WLN, chemical name, or chemical and physical data sheet for each compound.

To use this system, a controlled vocabulary is unnecessary. By using the menu screen and indicating the functions to be performed, the system allows searches, lookups, and file maintenance to be performed without having to memorize a procedure.

One of the facilities available to this online search system is the capability to save hit files. In order to save a search,

```
COMMAND: SEARCH ON SPECIFIC FRAGMENTS
SEARCH SUMMARY: (NC -R -02 AND G -R -04)

ENTER SEARCH CRITERIA BELOW:
AND/OR/NOT FRAGMENTS (SEPARATE WITH COMMAS FOR MINOR "OR")

AND          NC -R -01
              G -R -04

NUMBER OF COMPOUNDS: 00003
DISPLAY: LIST WLN      NAME      C&P
SAVE Y      HITS NAME  CHLOROTH

HITS FILE SAVED
```

Figure 5.

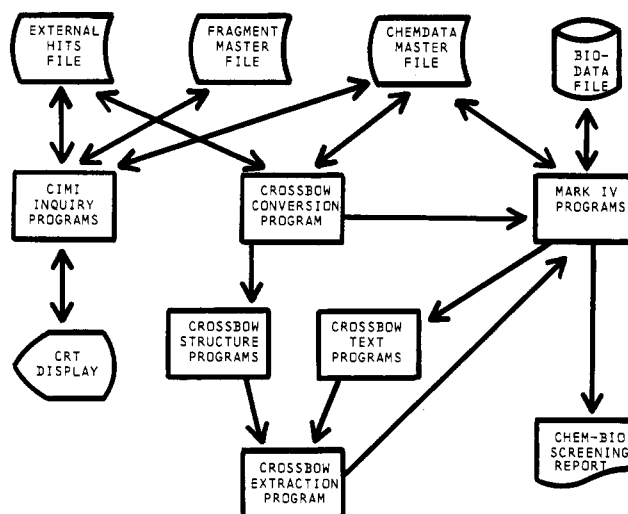


Figure 6.

as illustrated in Figure 5, the cursor is moved to SAVE, a Y is entered for "yes", and a descriptive name is entered for the hit file. The system responds with a message that the hit file has been saved.

If a hit file had been saved at an earlier session at the terminal, the hit file name can be ascertained by retrieving and looking at the catalog of external hit file (see Figure 4) so that it is unnecessary to remember the name by which a hit file has been saved. A hit file is saved not only for the one session at the terminal but will remain on file until such time that the user takes positive action to remove it from storage by exercising the command "Erase Hits". When a hit file from an earlier search is requested, a summary of the search criteria is also retrieved. A search saved on the external hit file may be written back to a current hit file for further refinement and perusal. Any search which can be performed against the master file can then be performed against a hit file to further refine it.

Whenever a function or operation of the system is not completely understood, a "Help" command will retrieve a set of instructions for all the functions of the system which have been implemented.

At the present time, the Diamond Shamrock chemical inquiry system is online on an IBM 3033. The Information Services group is equipped with an IBM 3276 control terminal, two satellite 3278 terminals, and a 3287 bidirectional printer which copies screen images to paper. Batch processing is initiated on the same IBM terminals using ROSCOE⁶ for time-sharing operations. We also have the option of initiating batch processing on the IBM 3033 by remote job entry (RJE), using a PDP 11/70 minicomputer and VT-100 terminals by Digital Equipment Corporation (DEC). Computer output reports are returned to a file on the PDP 11/70. They can

02/20/81

DIAMOND

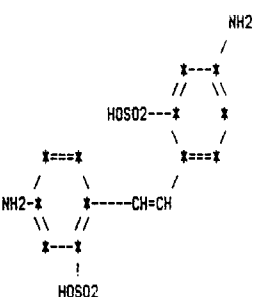
SHAMROCK

CORPORATION

2

DS 039237 BOT C14H14N2O6

WSQR CZ F1U1R DZ BSWQ



SOURCE BOT 01840

ACCEPTANCE DATE 08/01/78

SOURCE NAME PURCHASED

REF. NO.

SAMPLE SIZE 2 KG

CHEMICAL NAME 4,4'-DIAMINO-STILBENE-2,2'-DISULFONIC ACID

HERBICIDE WEEDS

TEST DATE	BATCH SUFFIX	TEST TYPE	DOSE KG/HA	PIG- WEED	VEL- LEAF	MUS- TARD	RED MIL	FOX- TAIL	B.Y. GRSS	JOHN GRSS	COF- FEE	MORN GLRY	WILD OATS	COMMENTS
09/05/78	0101	PRE	8.00	20	0	10	0	0	0					
09/05/78	0101	POST	8.00	30	10	0	0	10	0					

HERBICIDE CROPS

TEST DATE	BATCH SUFFIX	TEST TYPE	DOSE KG/HA	CORN	SOY- BEAN	WHE- AT	COT- TON	SUGR BEET	RICE	COMMENTS
09/05/78	0101	PRE	8.00	0	0		0			
09/05/78	0101	POST	8.00	0	0		0			

INSECTICIDE

TEST DATE	BATCH SUFFIX	DOSE PPM	BEAN BTLE	ARMY WORM	HSE- FLY	DOSE KG/HA	*** S+S	MITE SPR	*** SYS	*** S+S	APHID SPR	*** SYS	DOSE KG/15 CM HA	ROOT WORM	COMMENTS
09/22/78	0101	128.00	0	0	0	4.00	0			0			8.00	0	

PLANT HEALTH

TEST DATE	BATCH SUFFIX	DOSE PPM	DOSE KG/HA	***** FOLIAR PATHOGENS *****						***** SOIL PATHOGENS *****							
				BROWN S+S	SPOT SPR	SYS	BEAN S+S	RUST SPR	SYS	BEAN S+S	MILDEW SPR	ROTR CINE	RHIZ SOLA	PYTH APHA	THIE BASI	SCLR SCLR	ROOT KNOT
08/28/78	0101	1200.00											0				

IN VITRO ANTIMICROBIAL

EXPER. NUMBER	TEST DATE	BATCH SUFFIX	MIC TEST	DOSE PPM	***** E.COLI			***** S.CHOL			***** S.FAEC			***** BACTERIA C.PERF			***** S.AERU			***** P.AERU			***** P.HIRA		
79	09/19/78	0101		50.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0							

Figure 7.

then be viewed in 132-character format and subsequently submitted to a high speed line printer for paper copies.

Many chemists and biologists prefer to have searches performed for them by our staff of information chemists so that they themselves might use their time more profitably in the synthesis or biological laboratories. In response to that need, our most recent progress has been directed toward developing printed reports which include two-dimensional structural diagrams, chemical information, and complete biological screening data. These printed reports can be the result of batch searching or the interactive searching of our IMS online data base. To achieve this, we are using the CIMI online search programs interfaced with the CROSSBOW⁷ suite of search and structure generation programs and MARK IV Systems.⁸ The coordination of the files is best illustrated by outlining the process as shown in Figure 6. A search is performed on the online system and stored in the external hit file. A conversion program, written by Fraser Williams,⁹ retrieves the compound numbers from the external hit file and creates an input file to the CROSSBOW structure generation program. The same list of compound registry numbers is used by a MARK IV program to extract text data from the online IMS data base and output it to the CROSSBOW text program.

CROSSBOW programs place the structure and the text side by side as a heading for the report. MARK IV programs use the same compound numbers to extract biological screening data from the biological data file and to prepare and format the rest of the report.

A sample screening report in Figure 7 illustrates the report format, with two-dimensional structural diagram, chemical name, sources, and other text information, and screening data for each biological testing area. The data reported for the compound in Figure 7 do not represent actual percent controls for that compound. These chemical-biological screening reports are available on paper or on Computer Output Microfiche (COM).

We believe that these screening reports are an indication of our progress toward achieving the goal of integrating the results of chemical searches with biological screening data and making that data available to scientists in the Bioscience research areas.

REFERENCES AND NOTES

- (1) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation", 3rd ed.; Chemical Information Management, Inc.: Cherry Hill, NJ, 1975.

- (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.