# Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation*

R. J. FELDMAN and D. A. KONIVER**
Division of Computer Research and Technology,
National Institutes of Health,
Department of Health, Education and Welfare,
Bethesda, Md. 20014

Received April 21, 1971

**An interactive search and retrieval system for Wiswesser Line Notation (WLN) has been implemented. The system employs bit screens, which are useful for filtering a file. The user can graphically specify a search request structure and immediately receive graphic information as the result of the search. Four Fortran IV programs were developed to prepare bit screens for WLN files, input the search request to generate the WLN, iteratively search the WLN bit screen file, and generate a two-dimensional representation of the chemical structure directly from the WLN.**

In recent years, a number of chemical information systems have been developed for batch oriented computers. In these systems, the user codes a search request on a paper form. Cards are then punched from the paper form. The search requests are collected, and the search job is run on a batch computer. The user can then expect results after a delay of from one hour to two days. At N.I.H., we have been looking for ways of reducing the number of steps as well as the time delay between the asking of a question and the production of useful results. We have found that the chemist user of an information system is capable of conveying his request to the computer more adequately when his

*Presented before the Division of Chemical Literature, 160th Meeting, ACS, Chicago, Ill., Sept. 15, 1970.

**To whom correspondence should be addressed.

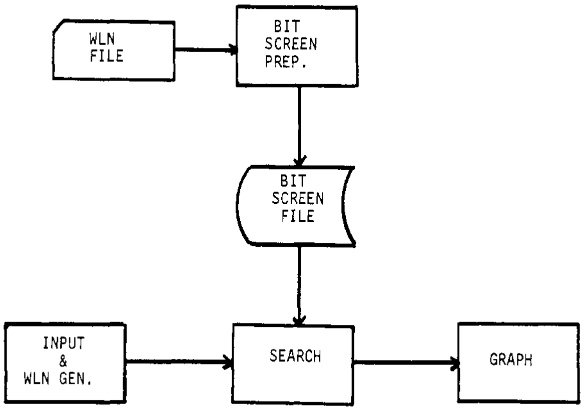
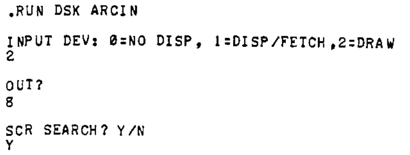
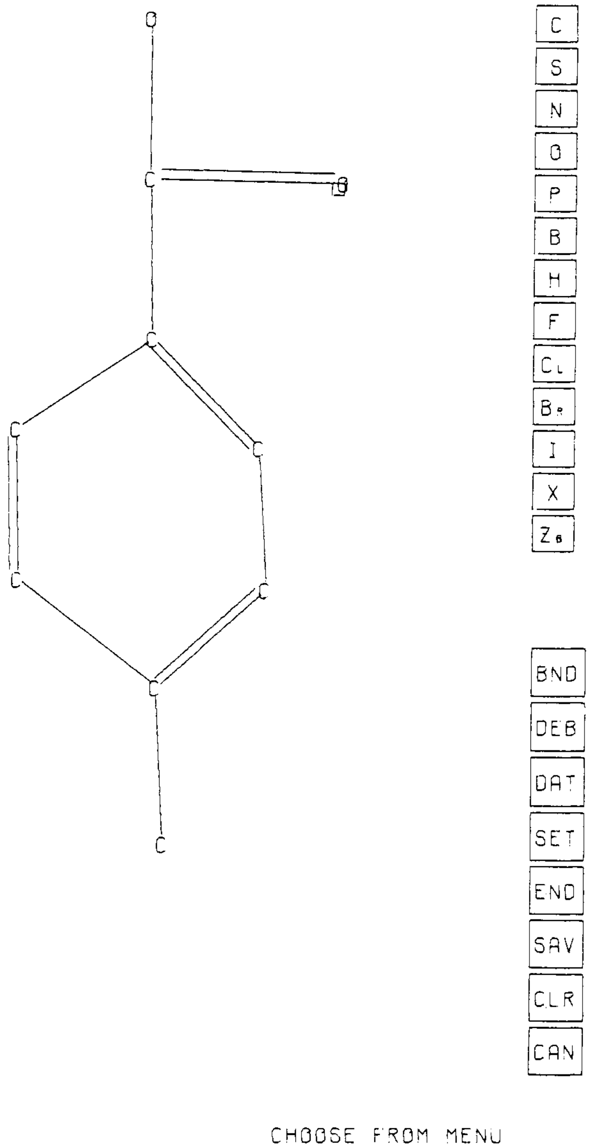


Figure 1. Program flow for the WLN search system



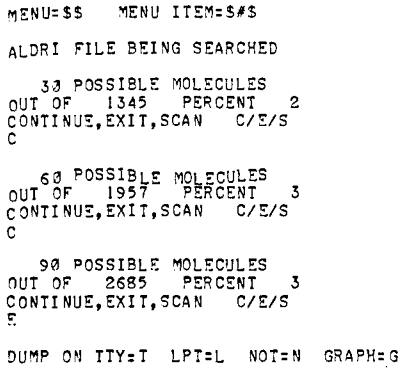
```
.RUN DSK ARCIN

INPUT DEV: Ø=NO DISP, 1=DISP/FETCH,2=DRAW
2

OUT?
8

SCR SEARCH? Y/N
Y
```

Figure 2. Initialization of the graphic input and WLN generation program



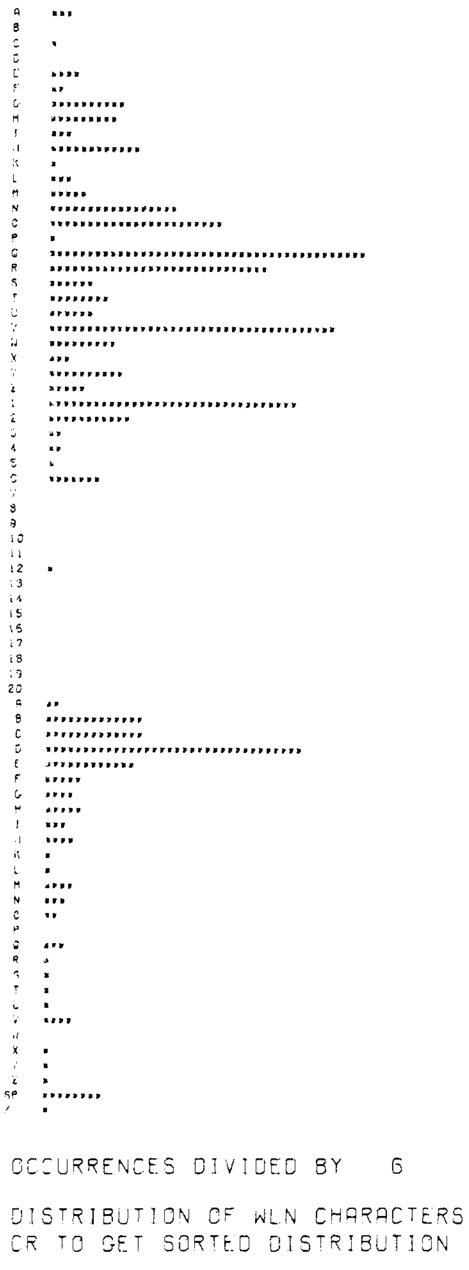
CHOOSE FROM MENU

Figure 3. Graph for the request "QVR D"



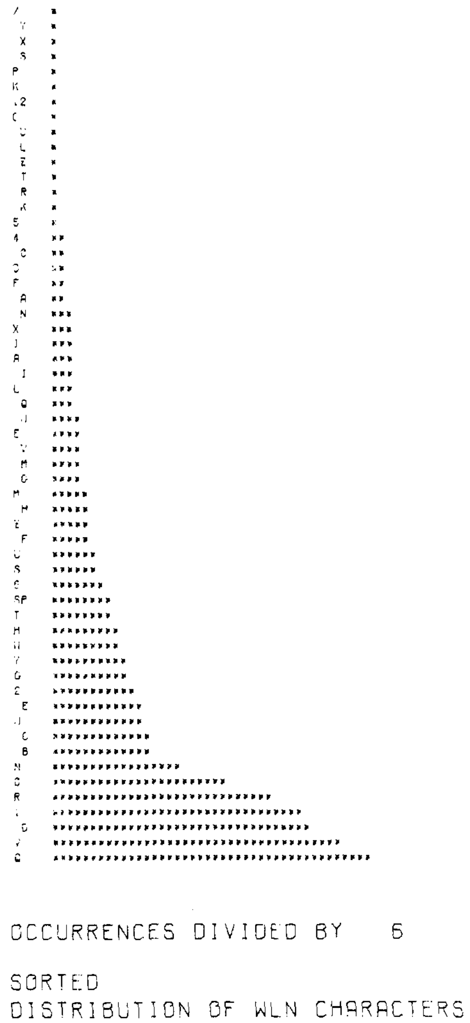
```
MENU=$$    MENU ITEM=$#$

ALDRI FILE BEING SEARCHED

   30 POSSIBLE MOLECULES
OUT OF    1345    PERCENT    2
CONTINUE,EXIT,SCAN    C/E/S
C

   60 POSSIBLE MOLECULES
OUT OF    1957    PERCENT    3
CONTINUE,EXIT,SCAN    C/E/S
C

   90 POSSIBLE MOLECULES
OUT OF    2685    PERCENT    3
CONTINUE,EXIT,SCAN    C/E/S
E

DUMP ON TTY=T  LPT=L  NOT=N  GRAPH=G
```

Figure 4. Teletype output for the first iteration of the search program



```
OCCURRENCES DIVIDED BY    6

DISTRIBUTION OF WLN CHARACTERS
CR TO GET SORTED DISTRIBUTION
```

Figure 5. Histogram of WLN syntactic units for the first iteration of the request "QVR D"



```
OCCURRENCES DIVIDED BY    6

SORTED
DISTRIBUTION OF WLN CHARACTERS
```

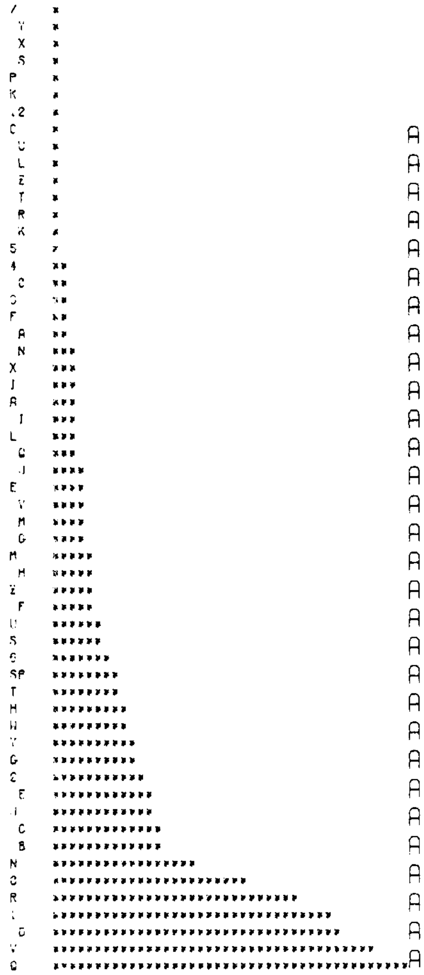
Figure 6. Sorted histogram of the WLN syntactic units resulting from the first iteration of the request "QVR D"

request is specified in graphic terms. Similarly, the results of a search also seem to be more useful to the chemist when they are presented to him in graphic terms.

A time-sharing computer, the PDP-10, has been used to implement just such an experimental search system, which permits graphic specification of search requests and graphic output of search results. During the search itself, the user is provided with graphic information which permits him to modify dynamically the search strategy at any point during the search. The immediate chemist-computer interaction eliminates the need for paper forms and punched cards.

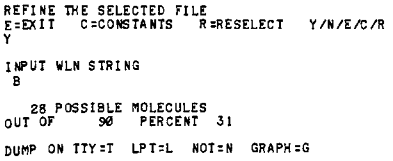
The system described in this paper is composed of four Fortran IV programs. The flow of these programs is given in Figure 1. This search system uses the Wiswesser Line Notation (WLN) as the basis for structure storage and search. The screen preparation program uses the bit screen notation of Granito *et al.*[2] The screen program is run whenever a new WLN file is obtained. The graphic input and WLN generation program was reported in a previous paper.[1] The search program uses the bit screen file as its data base input.

Requests to the search program can be entered from the graphic input and WLN generation program or from teletype input to the search program itself. Figure 2 shows the command to the PDP-10 to run the graph input and WLN generation program. The answer "Y" to the question "SCR SEARCH Y/N" means that, after the graphic request



```
INPUT WLN: QVR D


A 10053-6   ZR DVQ
A 10056-0   L6V DYJ B F1M1VQ DUYR BSWO&R DQ
A 10068-4   L6V DYJ BY C FY B1M1VQ DUYR BSWO
A 10087-0   QVYR DG&R DG
A 10098-6   L6V DYJ BE FE DUNR DQ C E1N1VQ1V
A 10204-0   QVR BG EG DVQ
A 10212-1   T56 BVO DHJ D- D-/R DQ CN1VO&1VO
A 10216-4   QVYZR CQ DQ
A 10233-4   QVR B E DVQ
A 10260-1   QV2R CQ DQ
A 10301-2   QV1U1R CQ DO1
A 10403-5   QV1MVR DR
A 10444-2   QVR B C E F DVQ
A 10786-7   ZMSWR DVQ
A 10808-1   1R DVOYVQ& 2
A 10851-0   QVR DE
A 11198-8   QR DVO2
A 11217-8   QR DV1
A 11284-4   QVYR D&R D
A 11340-9   WNR D BVQ
A 11371-9   1Y&R BQ D EV1
A 11428-6   QVV1R DQ
A 11483-9   1OR DYVQU1R DO1
A 11527-4   QVR BE DVQ
A 11533-9   T56 BMJ D2N1&1 GO1R &QVVQ
A 11739-0   QVR DO1
A 11743-9   L6V DYJ B F1N1&1VQ DUYR BSWO&R D
A 11886-9   QVXQR DF&R DF &QH
A 11969-5   QVR DM1
A 11976-8   WSQR DQ CNUNYR&UNMR BVQ
A 11983-0   T C666 BO EVJ DF FE IR BVQ& LE M
A 11988-1   WNR DNUNR DQ CVQ
A 11995-4   L C666 BV IVJ DQ GMVR& NMVR
A 11997-0   WNR CNUNR DQ CVO &-NA-
A 12003-0   OVR BQ DMVR &-CA-

OCCURRENCES DIVIDED BY    6

SORTED
DISTRIBUTION OF WLN CHARACTERS
```

**Figure 7. Histogram and WLN code resulting from the first iteration of the request "QVR D"**



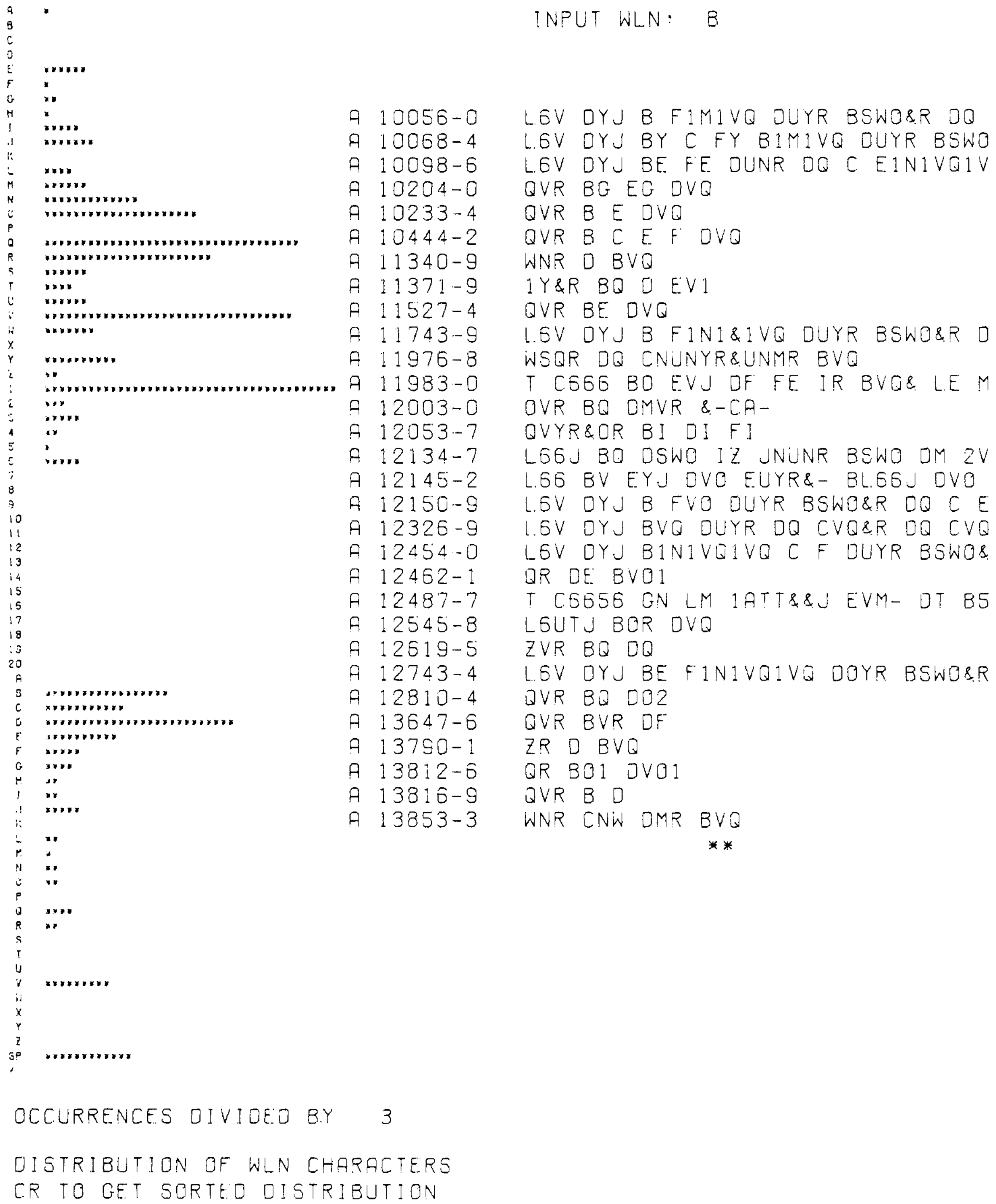
```
REFINE THE SELECTED FILE
E=EXIT   C=CONSTANTS   R=RESELECT   Y/N/E/C/R
Y

INPUT WLN STRING
B

  28 POSSIBLE MOLECULES
OUT OF     90   PERCENT  31

DUMP ON TTY=T  LPT=L  NOT=N  GRAPH=G
```

Figure 8. Teletype output for the second iteration of the search program

has been made, the program image for that program will be replaced by the program image for the search program.

Figure 3 shows the graphic request which causes the generation of the WLN code "QVR D". The WLN code is passed to the search program by the generation of a file on the system disk.

Figure 4 shows the teletype output of the search program as it starts. The Aldrich Chemical Company's WLN file of 8458 structures is being searched in this instance, although we also have a file from the Baker Chemical Company. The search program has a cutoff which can be set by the user. For this demonstration, the cutoff was set at 30. After 30 WLN matches on the file are found, the program asks if the user wishes to (1) continue generating WLN matches, (2) continue to scan the input WLN file but not record WLN matches, or (3) exit. In the case of Figure 4 the "C" (continue) command was given, and the program notified the user when another 30 matches were found. The user typed "C" again and the next time typed "E" (exit) knowing that 90 matches were found after searching 2685 WLN strings, or that about 3% of the file satisfied his request. A match is determined if a WLN string contains each of the symbols "Q", "V", "R", " D", although not necessarily contiguous to each other. Contiguity specifications are available and will be discussed later.



```
                              INPUT WLN:  B

              A 10056-0   L6V DYJ B F1M1VQ DUYR BSWO&R DQ
              A 10068-4   L6V DYJ BY C FY B1M1VQ DUYR BSWO
              A 10098-6   L6V DYJ BE FE DUNR DQ C E1N1VQ1V
              A 10204-0   QVR BG EG DVQ
              A 10233-4   QVR B E DVQ
              A 10444-2   QVR B C E F DVQ
              A 11340-9   WNR D BVQ
              A 11371-9   1Y&R BQ D EV1
              A 11527-4   QVR BE DVQ
              A 11743-9   L6V DYJ B F1N1&1VQ DUYR BSWO&R D
              A 11976-8   WSQR DQ CNUNYR&UNMR BVQ
              A 11983-0   T C666 BO EVJ DF FE IR BVQ& LE M
              A 12003-0   OVR BQ DMVR &-CA-
              A 12053-7   QVYR&OR BI DI FI
              A 12134-7   L66J BQ DSWO IZ JNUNR BSWO DM 2V
              A 12145-2   L66 BV EYJ DVO EUYR&- BL66J DVO
              A 12150-9   L6V DYJ B FVO DUYR BSWO&R DQ C E
              A 12326-9   L6V DYJ BVQ DUYR DQ CVQ&R DQ CVQ
              A 12454-0   L6V DYJ B1N1VQ1VQ C F DUYR BSWO&
              A 12462-1   QR DE BVO1
              A 12487-7   T C6656 GN LM 1ATT&&J EVM- DT B5
              A 12545-8   L6UTJ BOR DVQ
              A 12619-5   ZVR BQ DQ
              A 12743-4   L6V DYJ BE F1N1VQ1VQ DOYR BSWO&R
              A 12810-4   QVR BQ DO2
              A 13647-6   QVR BVR DF
              A 13790-1   ZR D BVQ
              A 13812-6   QR BO1 DVO1
              A 13816-9   QVR B D
              A 13853-3   WNR CNW DMR BVQ
                                        **

OCCURRENCES DIVIDED BY   3

DISTRIBUTION OF WLN CHARACTERS
CR TO GET SORTED DISTRIBUTION
```

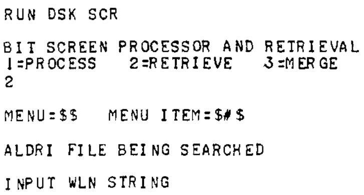
**Figure 9. Histogram and WLN code resulting from the second iteration**

The subfile of 90 WLN codes are placed in a file on the system disk. The program then asks on which medium the WLN matches should be outputted. "TTY" means teletype, "LPT" means line printer. In the case of Figure 4, the teletype was selected.

As the search of the input file is completed, the statistics of the WLN matches are presented on the display as a histogram. Each row of the histogram represents one WLN syntactic unit. Thus "R", "5", "19", and " B" are each WLN syntactic units. The occurrence of WLN units in the histogram in Figure 5 is in this case divided by six to scale the histogram picture to the size of the screen.

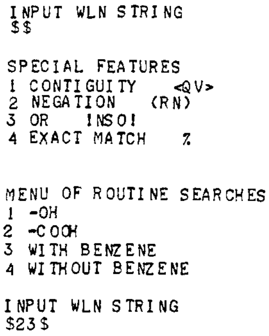
The histogram gives the user both a qualitative and a quantitative feeling for the distribution of WLN syntactic units of the selected subfile. In the case of the distribution of Figure 5, which resulted from the request of Figure 3, the largest rows in the histogram are for "Q", "V", "R", and "D". The largest rows correspond to the units in the initial request. The histogram also gives the user a feeling for the other WLN syntactic units in the subfile and thus structural components which have been included in the subfile as the result of partial matches. When an extra carriage return is given on the teletype, the distribution of WLN syntactic units is sorted and redisplayed, as in Figure 6.

When the display medium question on the teletype is answered as "T", and the display is in use because the



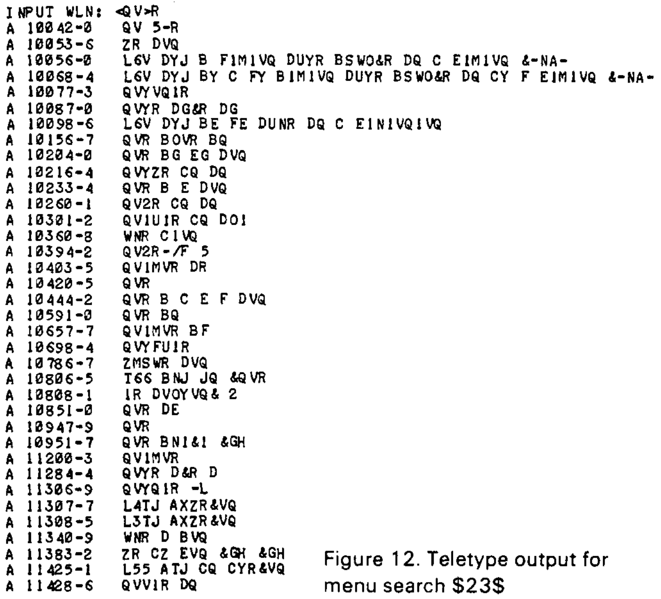
```
RUN DSK SCR

BIT SCREEN PROCESSOR AND RETRIEVAL
1=PROCESS   2=RETRIEVE   3=MERGE
2

MENU=$$    MENU ITEM=$#$

ALDRI FILE BEING SEARCHED

INPUT WLN STRING
```

Figure 10. Initialization of the search program for teletype input



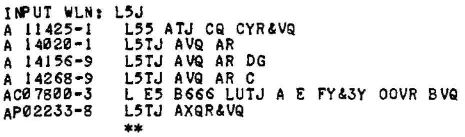
```
INPUT WLN STRING
$$

SPECIAL FEATURES
1 CONTIGUITY    <QV>
2 NEGATION   (RN)
3 OR   1NSO!
4 EXACT MATCH    %

MENU OF ROUTINE SEARCHES
1 -OH
2 -COOH
3 WITH BENZENE
4 WITHOUT BENZENE

INPUT WLN STRING
$23$
```

Figure 11. Teletype output of the special feature and menu search lists



```
INPUT WLN: <QV>R
A 10042-Ø   QV 5-R
A 10053-6   ZR DVQ
A 10056-Ø   L6V DYJ B F1M1VQ DUYR BSWO&R DQ C E1M1VQ &-NA-
A 10068-4   L6V DYJ BY C FY B1M1VQ DUYR BSWO&R DQ CY F E1M1VQ &-NA-
A 10077-3   QVYVQ1R
A 10087-Ø   QVYR DG&R DG
A 10098-6   L6V DYJ BE FE DUNR DQ C E1N1VQ1VQ
A 10156-7   QVR BOVR BQ
A 10204-Ø   QVR BG EG DVQ
A 10216-4   QVYZR CQ DQ
A 10233-4   QVR B E DVQ
A 10260-1   QV2R CQ DQ
A 10301-2   QV1U1R CQ DO1
A 10360-8   WNR C1VQ
A 10394-2   QV2R-/F 5
A 10403-5   QV1MVR DR
A 10420-5   QVR
A 10444-2   QVR B C E F DVQ
A 10591-Ø   QVR BQ
A 10657-7   QV1MVR BF
A 10698-4   QVYFU1R
A 10786-7   ZMSWR DVQ
A 10806-5   T66 BNJ JQ &QVR
A 10808-1   1R DVOYVQ& 2
A 10851-Ø   QVR DE
A 10947-9   QVR
A 10951-7   QVR BN1&1 &GH
A 11200-3   QV1MVR
A 11284-4   QVYR D&R D
A 11306-9   QVYQ1R -L
A 11307-7   L4TJ AXZR&VQ
A 11308-5   L3TJ AXZR&VQ
A 11340-9   WNR D BVQ
A 11383-2   ZR CZ EVQ &GH &GH
A 11425-1   L55 ATJ CQ CYR&VQ
A 11428-6   QVV1R DQ
```

Figure 12. Teletype output for menu search $23$



```
INPUT WLN: L5J
A 11425-1   L55 ATJ CQ CYR&VQ
A 14020-1   L5TJ AVQ AR
A 14156-9   L5TJ AVQ AR DG
A 14268-9   L5TJ AVQ AR C
ACØ7800-3   L E5 B666 LUTJ A E FY&3Y OOVR BVQ
APØ2233-8   L5TJ AXQR&VQ
            **
```

Figure 13. Teletype output of subfile satisfying the requests "⟨QV⟩R" and "L5J"

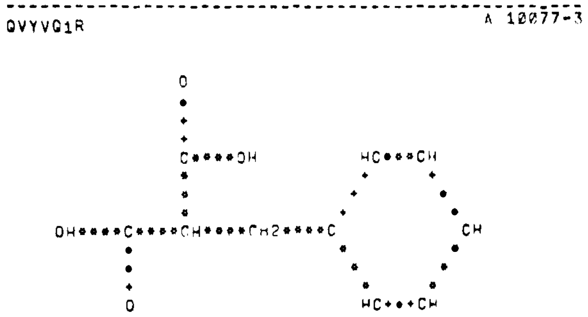


Figure 14. A graph from the subfile of "⟨QV⟩R"

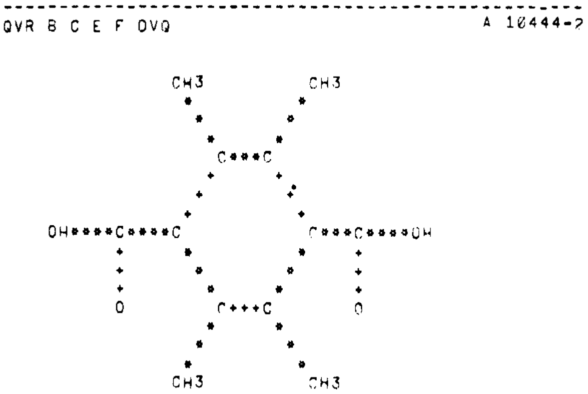


Figure 15. A graph from the subfile of "⟨QV⟩R"

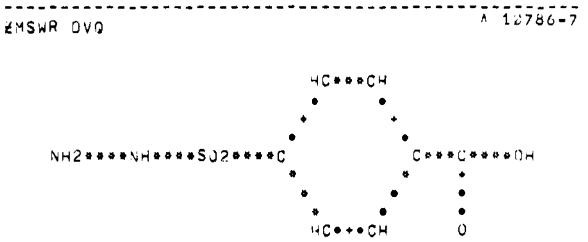


Figure 16. A graph from the subfile of "⟨QV⟩R"

initial search request came from the graphic input and WLN generation program, a portion of the selected subfile is presented on the display, along with the sorted histogram. In Figure 7, the serial number on the left side of the WLN code is the Aldrich catalog number.

When the user is finished with the first iteration of the search, the user gives the response "Y" on the teletype to the question "REFINE THE SELECTED FILE Y/N". The search program then requests the user to input a WLN string to further refine the selected subfile. In the case of Figure 8, the user responded with " B". This causes the search program to record in the next subfile only those structures with a " B" locant. After 30 structures are found, the program again asks if it should continue. In the case of Figure 8, the user responds with "E". The program tells the user that 31% of the subfile searched so far meets the requirements (cumulatively) of "QVR D B".

Figure 9 is analogous to Figure 7. The distribution of WLN syntactic units has been modified by forcing the presence of " B". In Figure 9, the histogram is unsorted.

The iterative process of refining the selected subfile proceeds very rapidly. The search time for the subfile gets progressively smaller. The process instead of being computer limited becomes human response limited. At any point in the iteration the user has enough information to make a relevant decision to constrain further the selected subfile.

The search program can be used as the only source of search request. When operating in this mode, the display is not used, and all output goes either to the teletype or to the line printer. Figure 10 shows the initialization of the search program in this mode. Capability exists for (1) generation of WLN bit screens, (2) normal search and retrieval, and (3) merging of WLN bit screen files. The special features can be obtained by typing "$$" as the input WLN string.

The special features of the search program permit the user to perform any of four logical actions during the search. By using the contiguity feature, the user can force at least one occurrence of the syntactic units between the angle brackets to be contiguous. The WLN codes "QR BVI" and "QR BVQ C2" would both satisfy the request "QVR" but only "QR BVQ C2" would satisfy the contiguity of "⟨QV⟩R". By using the negation feature, the user can inhibit the presence of the WLN syntactic units between the parentheses. The negation feature is very difficult to use since all occurrences of the negated syntactic units cause the rejection of a structure. The OR feature permits a structure to be selected if the WLN code contains any of the syntactic units between the exclamation marks. An exact match between the requested structure and any selected structure is forced by preceding the teletype request string by a "%".

The menu of standard searches permits the user to select easily certain broad subfiles. The input string "$23$" for a menu search would be translated into the WLN search request "⟨QV⟩R". Figure 12 shows a partial listing of the result of the search on the Aldrich file.

The subfile generated by the above menu search is further constrained by specifying that all structures must have a five membered carbocyclic ring. Figure 13 shows the result of the request.

When the user has iteratively selected a subfile which has the required structural properties, the two dimensional graph of the structures can be generated by typing "G" in response to the output medium question (Figure 8). The program image of the search program is replaced by the graph program image. The most recently selected subfile is used as the input to the graph program. The graph program works directly from WLN, making use of information such as the fact that at a branch point the shortest chains will usually be cited first.

Figures 14 through 18 are typical two dimensional graphs generated by this program. The graph program represents a single bond by a string of stars (*****), a double bond by a string of plusses (++++++), and a triple bond by a string of number marks (#####). Figure 17 rather clearly illustrates one of the artifacts of this program. Whenever a long arm occurs on the B or F locants of a ring, the arm is bent so that the rest of the graph is generated from left to right rather than from top to bottom. Figure 18 illustrates some of the problems still to be solved in this program. When the structure is complex, portions of the representation tend to overlap.

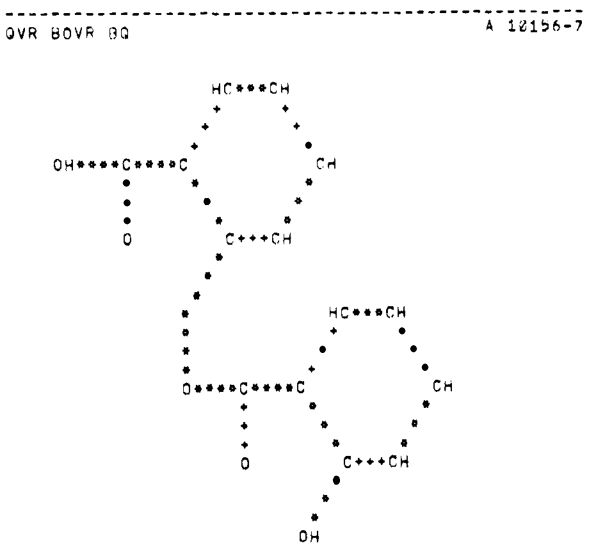
The two dimensional graphs of molecular structures can be outputted to either the teletype or the line printer. The program produces structure representations at line printer speed.

## CONCLUSION

Graphic input and output of chemical information as developed in this search and retrieval system brings the chemist user closer to the computer by adapting the computer to the methods of thought and representation of the user. The interaction of the user with the search strategy permits the user to shorten the time for feedback cycle to a matter of seconds.

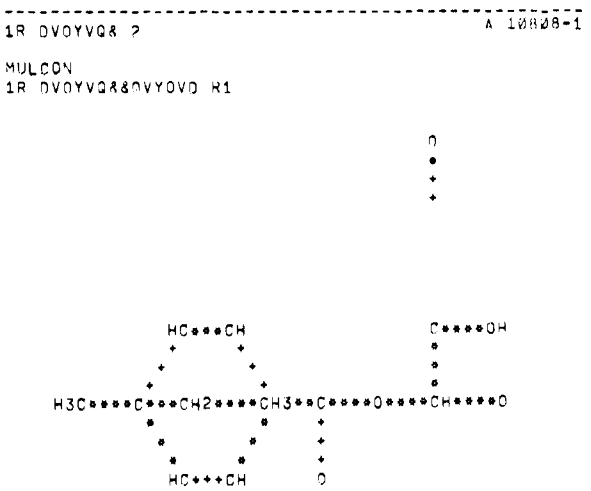
The experience gained in the development of this experimental search and retrieval system has led to significant modification of our view of data base management, searching, and graphic structure representation.

When a structure is coded in WLN, contiguous patterns of atoms tend to become separated. This is an artifact of the linearization of the topology of the structure. Searching for contiguous patterns of atoms encoded in WLN is very difficult. We are currently looking into methods of doing substructure search from connection tables.



```
QVR BOVR BQ                                        A 10156-7
```

Figure 17. A graph from the subfile of "⟨QV⟩R"



```
1R DVOYVQ& ?                                       A 10808-1

MULCON
1R DVOYVQ&&OVYOVD R1
```

Figure 18. A graph from the subfile of "⟨QV⟩R"

The WLN syntax requires a complicated program to do graphic structure representation. The graph program discussed in this paper can handle up to fused rings, but has great difficulty with multipliers, bridged rings, and spiros. The results of this program, while supporting the feasibility of structure generation from WLN, admittedly do not compare with the excellent output from I.C.I.'s Crossbow Program.[3] We are currently looking at methods of graph representation from connection tables. The direct representation process does away with the necessity of coping with WLN syntax per se.

## LITERATURE CITED

(1) Farrell, C. D., A. R. Chauvenet, and D. A. Koniver, "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.* **11**, 52–9 (1971).

(2) Granito, C. E., G. T. Becker, W. J. Wiswesser, and K. J. Windlinx, "Computer-Generated Superimposed Codes for Searching Chemical Substructure Files," *J. Chem. Doc.* **11**, 106–10 (1971).

(3) Rogers, M.A.T., "Crossbow," paper presented at the 158th Meeting, ACS, New York, N. Y., September 1969.