GRAPHIC DISPLAY OF THE STRUCTURAL COMPOSITION OF CHEMICAL COMPOUNDS

W. M. Verbestel
Computer Systems Directorate
Systems Planning and Research
Revenue Canada, Customs & Excise
270 Albert Street,
Ottawa, Ontario
K1A 0L5

C.Y. Suen
Department of Computer Science
Concordia University
1455 de Maisonneuve West
Montreal, Quebec
H3G 1M3

ABSTRACT

Displaying the structure of a chemical compound in graphical form, as the end point of a Chemical Information System, is essential if mechanization is to be acceptable to the organic chemist. In practice, from the chemists viewpoint, any system proposed for Data Processing will be incomplete without this facility, and the quality of the output must be consistent with the state of the art in this area. This paper gives details of a computer program for the regeneration of organic chemical structures derived from the Wiswesser notation.

La représentation graphique de la structure des composés chimiques, comme point final d'un système d'information, est essentiel si la mécanisation doit être acceptable au chimiste organicien. Du point de vue du chimiste, tous systèmes proposés pour le traitement de données sera incomplet sans cette facilité, de plus la qualité du graphique obtenu doit être à la pointe du progres dans ce domaine. Cet article donne des détails d'un programme d'ordinateur régénerant des structures de chimie organique dérivant de notation Wiswesser.

INTRODUCTION

The chemical literature is enormously large. Many different kinds of chemical compounds exist and many new ones are discovered every day. Hence, the information chemists whose job it is to search the chemical literature, are faced with the impossible task of searching tens of thousands of compounds, and therefore must rely to an increasing extent on the support of automated systems. Even with the aid of a computer, these chemical compounds must be encoded, preferably in a linear form, to facilitate the creation of a data-bank, and to assist in the subsequent retrieval process [1]. One technique commonly used is the Wiswesser Line Notation (WLN) which encodes the chemical compound using a series of alphanumeric and special symbols [1,2]. While WLN can give a compact linear form for chemical compounds, it unfortunately does not give their structural composition - which is the most important information required by the chemist. Furthermore, it is very difficult for a chemist to decode the WLN manually due to the complex rules used in the encoding process. In view of the above, a software package has been developed to decode the WLN and display on a graphic terminal the structural form of the chemical code entered. This display package was written in Fortran IV and implemented on a XEROX Sigma 9 computer with a Tektronix 4015 display terminal.

The Wiswesser Line Notation

The Wiswesser Line Notation (WLN) [2] is a precise and concise means of expressing the structural formula of chemical compounds. It uses letter symbols to denote functional groups, and numbers to express the lengths of alkyl chains and the sizes of rings. These symbols are specified in sequential order from one end of the molecule to the other, and the 40 symbols used employ the ten numerals, 26 capital letters, three special characters (&, - and /), and the blank. Alternate starting points and the choice of alternate paths through the structure, are governed by the position of the symbol in the alphabet.

A short list of WLN symbols follow:

- All the international atomic symbols are used except K, U, V, W, Y, Cl, and Br.
- Two letter atomic symbols in organic notation are placed between hyphens.
- Single letters preceded by a blank space indicate ring positions.
- Numerals preceded by a space are multipliers of preceding notation symbols, or, within ring symbols (L... J or T... J) show the number of multicyclic points in the ring structure.
- Numerals not preceded by a space show ring sizes (if within the ring signs) - elsewhere, numerals show the length of internally saturated, unbranched alkyl chains and segments.
- Single letters not preceded by a blank space have the following meaning:
  - A, B, E, F, G, I, J, K, M, N, O, P, Q, S, V, W, Z correspond to functional group (e.g. -CO-) or atoms.
  - L... J or T... J are used as enclosing symbols for a carbocyclic or heterocyclic ring notation.
  - X or Y are used as an alkyl branch symbol.
  - &, punctuation mark.

Philosophy of the Program

The program is composed of 5 modules viz:

1. The input and mapping module - which maps the input string of characters into a string of integer numbers to facilitate the decoding process and to detect any illegal Wiswesser character,
2. The ring decoding module - which detects the boundary of a ring code in the WLN, the ring size, the bridge, the unsaturations and the spiro locations. It also detects the locants of the ring, the multicyclic atoms in the molecule and the perifused atoms.

   This module converts the information received from the WLN into a connecting table, where an entry is provided for each element considered as a node in the display. The table is 20 words wide and consists of:

Position	Description
1	Internal pointer showing the size of the table.
2 - 3	X and Y coordinates for this node (filled during execution of the conversion module).
4 - 11	4 two-words entries, a pointer to the node to be linked, and its angle (in degree).
12	counter showing the degree of saturation.
13 - 14	internal counters used to compute bond direction.
15	number of characters to be represented at the node.
16	pointer to the character stack.
17	multicyclic flag.
18	multiplier flag,
19	chain flag.
20	spiro flag.

3. A chain decoding module - which decodes a chain of elements (or chains of substituents) on a previously defined ring notation, and provides information on the same connecting table as described above.
4. A routine to carry conversions from the connecting table to a drawing table - which provides the absolute coordinates on the screen for each of the nodes in the connecting table, and creates a drawing table giving the absolute coordinates of the starting and the ending points for each vector to be drawn.
5. A structure display module - which develops a drawing factor to frame the virtual window for aesthetical display purposes on the screen. This module creates the screen graphic display from graphic segment coordinates stored in the drawing table and from a string of characters and their coordinates stored in the connecting table.

Program operation is archieved by having the monitor (the main routine) call the input module, after which the latter will sequentially scan the WLN string and call the ring decoder or

the chain decoder as appropriate. At the conclusion of this execution, the monitor will call the conversion and drawing modules.

These last two modules are concerned with the plotting of each atom and bond according to a free plotting routine. The conversion routine actually computes the (X,Y) coordinates required to position each atom and its associated bonds correctly.

Positional linking of atomic groups is a specific case of linear plotting, but directional changes become necessary when a branching unit is approached. The program must derive the appropriate angles (or "new directions") of track to plot atoms and bonds of side branches, & ring atoms also require special consideration. The routine must compute the coordinates for ring atoms and bonds to give a closed ring diagam of regular polygonal figures. Complexities in plotting arise in attempting to overcome any overwriting errors, as the program must not only recognize a situation where an overwriting conflict is likely to occur, but must then have the ability to modify its tracking route to avoid the error. As well as having the ability to plot a structure without errors, the display program must be able to construct an image which will be readely recognisable to the eyes of a chemist.

A large set of error messages are coded under the supervision of the monitor, and each module may return an error code to the monitor. Typically, errors messages can be generated as a result of invalid WLN characters, invalid WLN codes or code not implemented within the graphic display (e.g. Kelate, some kind of complex bridges, AminoAcid contaction codes etc...).

Comparison of this graphic system with others published in the Chemical Literature

This system is coded with a view to its being used at the end of a complete Chemical Information system based on the WLN Data Base. The graphic output module was created for a Tektronix on-line graphic console. However, as the core of the graphics system presents the output module with a stack of XY coordinates of the starting and ending points of segments to be drawn, the later can be rewritten for any kind of high resolution graphic device i.e. a plotter, a dot matrix printer or any kind of graphic terminals with at least 1000 x 1000 dot resolution.

The following papers were identified through a computer search of the chemical literature, and are concerned with printing devices used in the representation of chemical structure derived from WLN codes:

Rogers [3] employed a method of expanding WLN codes, creating a connecting table called CROSSBOW and displaying the structure using a modified print chain. Outputs are not compatible with current organic chemistry practices.

Carhart [4] and Feldman [5] used a teletype to represent the chemical structure through the use of a modified print element.

A better representation was obtained by Cottardi [6] but he was primarely concerned with the graphic device, not the WLN decoding. An effective and logical methodology was presented by Hyde [7] in his "Structure Display", however, there appears to have been little more specific work conducted in this area since 1975.

CONCLUSION

This computer program was developed in 1977 (as a thesis requirement for a Masters in Computer Science) by W. M. Verbestel under the supervision of C. Y. Suen at Concordia University. While it is considered that the proposed methodology has considerable potential, it is also recognised that changes may be necessary to accommodate changes in WLN standards.

References

[1] Index Chemicus Registry System Institute for Scientific Information Philadelphia, Pensylvania 19106.

[2] Smith,E.G. "The Wiswesser Line. Formula Chemical Notation", McGraw - Hill, New York, 1968.

[3] Rogers, M.A.T. "Organic Chemical Structures on Computers", Chem. and Ind., 18 July 1970, pp. 952-5.

[4] Carhart, R.E., "A Model Based Approach to the Teletype Printing of Chemical Structures", J. of Chem. Inf. & Comp. Sc., Vol 16, 2, 1976 pp. 82-8.

[5] Feldman, A., " Chemical Teletype", J. of Chem. Doc., Vol 13, 2, 1973 pp. 53-56.

[6] Cottardi, R. "A Modified Dot-Bond Structural Formula Font with Improved Stereochemical Notation Abilities", J. of Chem. Doc., Vol 10, 2, 1970 pp. 75-81

[7] Hyde, E. & Al, "Structure Display", J. of Chem. Doc., Vol 8, 3, 1968, pp. 138-145

WLN CODE
Z1YUGM1UMY1UG1UQ

CPU TIME SPENT = .46 SEC.

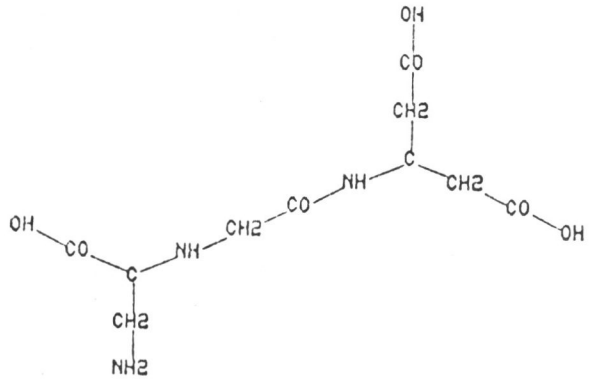

Fig. 1 Branching Chain

WLN CODE
ZR CYR CE&3MUR

CPU TIME SPENT = .81 SEC.

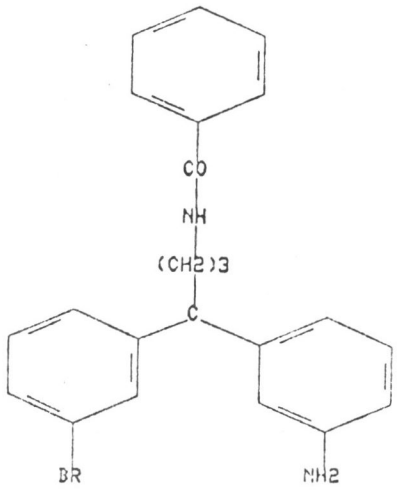

Fig. 2 Chain of Benzene Rings

WLN CODE
T5M CN BUTJ B1NR&1R

CPU TIME SPENT = .46 SEC.

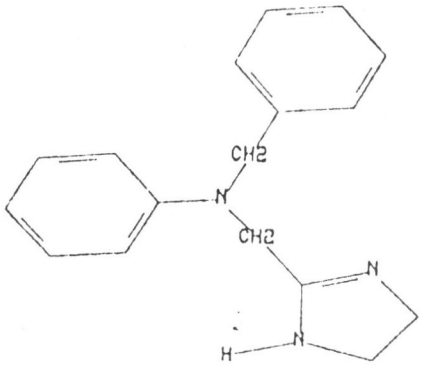

Fig. 3 Benzene and other Rings

WLN CODE
T D6 B656 LMJ CQ F01 G01

CPU TIME SPENT = .81 SEC.

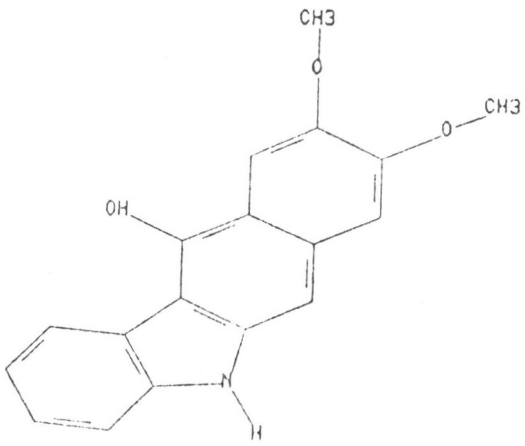

Fig. 4 Polyfused Substitued

WLN CODE
T F5 D6 B656 CN GO IO OM HHJ

CPU TIME SPENT = .82 SEC.

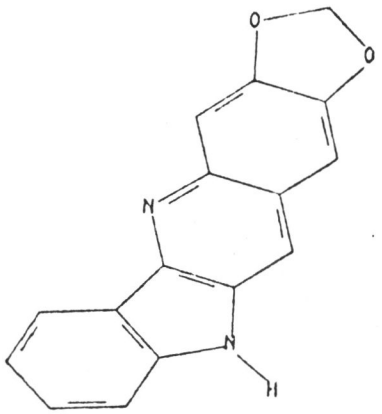

Fig. 5 Polyfused Heteroelements

WLN CODE
L E5 B666 MUTJ A E FV1Q FQ OQ

CPU TIME SPENT = .85 SEC.

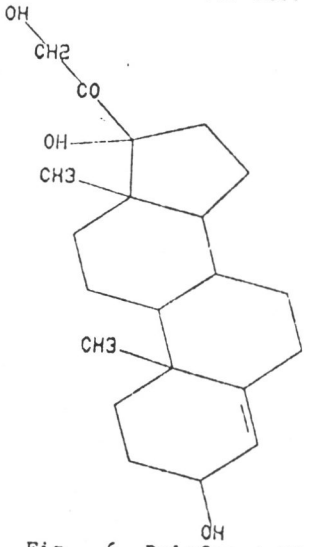

Fig. 6 Polyfused Rings

WLN CODE
T B6 H676 COJ

CPU TIME SPENT = .44 SEC.

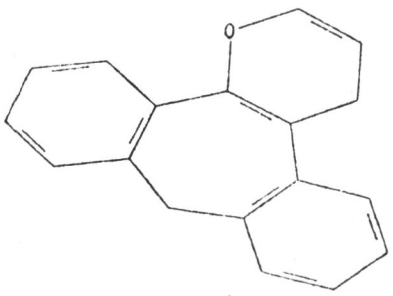

Fig. 7 Polyfused Rings

WLN CODE
T5NYMV EHJ A1 BUM E1R

CPU TIME SPENT = .44 SEC.

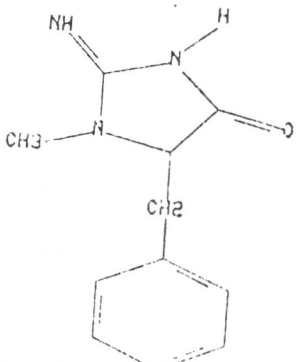

Fig. 8 Cyclics Rings

ULN CODE
T6NJ BN2N1&1&1- BT5SJ EE

CPU TIME SPENT = .80 SEC.

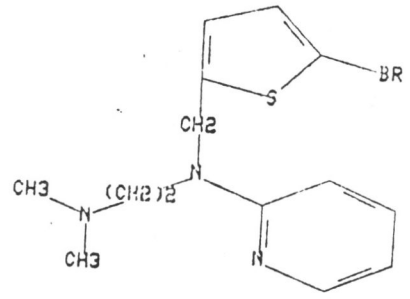

Fig. 10 Branching Chains of Heterocycles

ULN CODE
T6NJ BO2N2&2 FXQR DR&&- BT5OJ

CPU TIME SPENT = .89 SEC.

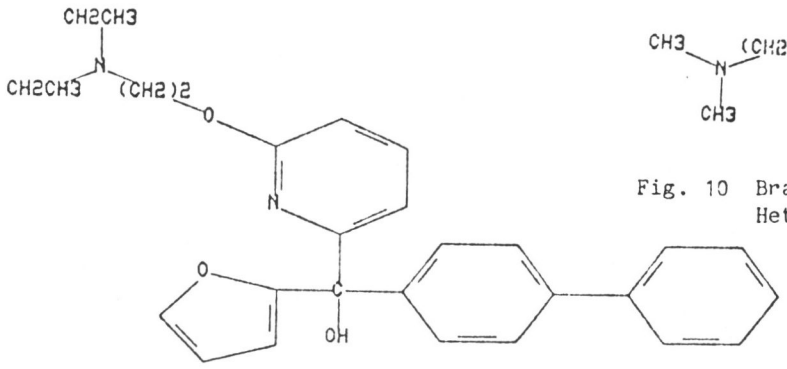

Fig. 9 Chains of Polyfused Rings

ULN CODE
T66 CNJ H- HT66 CNJ D- AL6TJ C- AL6TJ

CPU TIME SPENT = .86 SEC.

ULN CODE
L666 B6 C6 3ABC S EHJ

CPU TIME SPENT = .49 SEC.

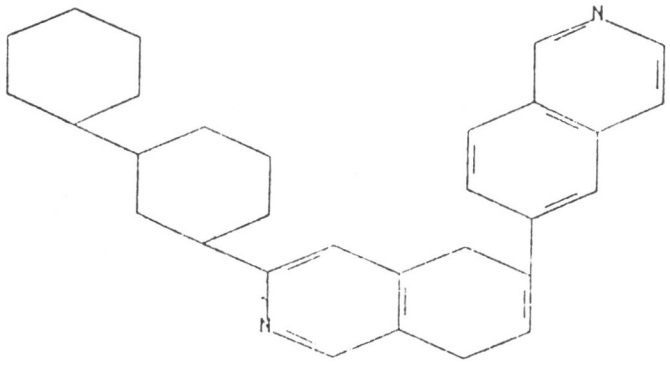

Fig. 11 Chains of Heterocycles

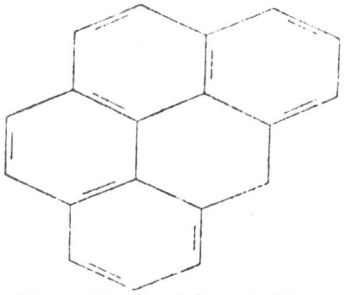

Fig. 12 Perifused Rings