N. Y., 1971, pp 2-263.

(2) Hansch, C., "Quantitative Approaches to Pharmacological Structure-Activity Relationships" in "Structure-Activity Relationships," Vol. I, C. J. Cavallito, Ed., Pergamon Press, Oxford, England, 1973, pp 75-165.

(3) Bruice, T. C., Kharasch, N., and Winzler, R. J., "A Correlation of Thyroxine-Like Activity and Chemical Structure," *Arch. Biochem. Biophys.*, **62**, 305-317 (1936).

(4) Free, S. M., Jr., and Wilson, J. W., "A Mathematical Contribution to Structure-Activity Studies," *J. Med. Chem.*, **7**, 395-399 (1970).

(5) Craig, P. N., "Comparison of the Hansch and Free-Wilson Approaches to Structure-Activity Correlation," *Advan. Chem. Ser.*, No. 114, 115-129 (1972).

(6) Verloop, A., "The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies" in "Drug Design," Vol. III, E. J. Ariëns, Ed., Academic Press, New York, N. Y., 1972, pp 133-187.

(7) Kier, L. B., "Molecular Orbital Theory in Drug Research," Academic Press, New York, N. Y., 1971.

(8) Reference 1, p 24.

(9) Colombo, D. S., and Rush, J. E., "Use of Word Fragments in Computer-Based Retrieval Systems," *J. Chem. Doc.*, **9**, 47-50 (1969).

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

(11) 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**, 106-110 (1971).

(12) Bird, A. E., and Marshall, A. C., "Correlation of Serum Binding of Penicillins with Partition Coefficients," *Biochem. Pharmacol.*, **16**, 2275-2290 (1967).

(13) A computer-sorted listing, updated semiannually, of log *P* values and electronic and steric parameters for over 800 substituents is available. Inquiries may be addressed to: A. Leo, Pomona College Medicinal Chemistry Project, Department of Chemistry, Pomona College, Claremont, Calif. 91711.

(14) Farrar, D. E., and Glauber, R. R., "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econom. Statistics*, **49**, 92-107 (1967).

(15) Haitovsky, Y., "Multicollinearity in Regression Analysis: Comment," *Rev. Econom. Statistics*, **51**, 486-489 (1969).

Computerized Management of Structure-Activity Data. II. Decoding and Searching Branching Chains and Multiplied Groups Coded in WLN

A. LEO,* DAVID ELKINS,† and CORWIN HANSCH

Department of Chemistry, Pomona College, Claremont, California 91711

Received December 21, 1973

As each WLN symbol for a structure containing a branching chain and/or multiplied groups is extracted in a left-to-right scan, the symbol to which it was connected in the *graphic* formula[1] must be known. For highly branched structures, especially where ring systems and/or multipliers are present, the program logic becomes quite complex. Ring substituents are discussed but decoding of the ring structure itself is reserved for the following article.

The increased use of Wiswesser Line Notation (WLN)[2] for the computerized storage, sorting, and searching of chemical structures[3-5] has created the need for programs that will either generate the notation from electronically drawn diagrams[6,7] or check the accuracy of manual encoding.[8] In the latter case the algorithms needed to properly dissect the WLN and compare it to a molformula are also useful to provide information which makes subsequent sorting and structure searching more efficient. The following WLN decoding programs are possibly similar to those in use elsewhere, but the particular algorithms for handling branched chains, multiplied groups, and rings have not been widely publicized, and they may prove helpful to those contemplating their own computerized structure files.

DISSECTION

The first step in the decoding process is the resolution of the various components of the WLN. The machine must be programmed to extract locants, multipliers, ring kernels (the kernel is that portion of a ring system set off by the ring initiator, D, L, or T, and the closing J), alkyl chain numerals, two-letter atom symbols (-NA-, -FE-), etc., from the string of simple structural symbols which comprises the essence of the WLN. This is accomplished by examining each character in its context during a left-to-right scan of the notation. When locants or ring systems are encountered, special routines are called upon to extract these items.

ATOM EQUIVALENCE

The next step is to convert the extracted WLN symbol into its atomic equivalent. Some WLN and atomic symbols are identical, such as I, O, and S, while others, such as M, V, and Z, denote multiatom fragments. Except for ring kernels which are handled by special routines (see the following paper in this series), conversion is made *via* a Symbol Equivalence, Table I.

Note that methyl-contracting branches (K, X, and Y), when they are not within rings, are entered with their full complement of methyl groups. Ring positions which are not specified as locations for heteroatoms or V, X, or Y are considered initially to be fully substituted with hydrogen atoms. This includes the benzene symbol R. An alkyl chain numeral, *n*, is entered as C_nH_{2n+2}.

VALENCE AND BOND FORMATION

The valence of each symbol (*e.g.*, the number of single chemical bonds available for attachment to other symbols) is determined from Table II.

* To whom correspondence should be addressed.

† Present address: G. D. Searle & Co., Chicago, Ill. 60680.

Table I. Symbol Equivalence[a]

WLN	When outside ring	When inside ring (if different)
E	Br_1	
G	Cl_1	
K	$C_3H_9N_1$	N_1
-KA-	K_1	
M	H_1N_1	
Q	H_1O_1	
R	C_6H_6	
W	O_2	
X	C_4H_9	C_1
Y	C_3H_7	C_1
Z	H_2N_1	

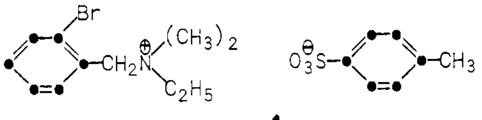
[a] For WLN symbols which differ from atomic symbols.

Table II. Valence of Symbols

Valence[a]	WLN symbol
1	E, F, G, H, I, Q, Z, -NA-
2	M, O, S,[b] V, W, -CU-
3	B, N, Y[a]
4	C, K, S,[c] U, X
5	P
6	UU

[a] The valence of certain symbols is altered by their context; the valence of an alkyl numeral is decreased to 1 when it initiates or ends the notation; it is increased by 2 when adjacent to UU; a Y or alkyl numeral's valence is increased by 1 when adjacent to a U. [b] Sulfur's normal valence. [c] Sulfur has a valence of 4 when adjacent to W, ZW, or terminal O, *i.e.*, an O which begins or ends the notation or ends a branch.

To reconcile WLN with molformula and to derive the correct atom connectivity, the sites of ionic charges must be specified at the end of the WLN (following a spaced ampersand).[9] The charges are then added to the appropriate valences. For example, in the WLN



1

ER B1K2 &WSO&R D &6/12

there is a positive charge on the K and a negative charge on the O. Therefore, K will have a valence of 4 + 1 = 5, and O will have a valence of 2 − 1 = 1. A preliminary scan of the notation determines the positions and magnitudes of the charges, and this information is utilized during the subsequent decoding scan to assign true valences.

THE TIP LIST

If we visualize the decoding process as the progression of a "growing tip," we can imagine it "passing through" one or more branching points, leaving "buds" where further branches are to be developed. As each WLN symbol is fully bonded, it is "scratched" from the Tip List and any symbols which follow are grafted on the immediately preceding "bud." Obviously, if the construction of a connection table (CT) is desired, the "tips" will be entered there rather than discarded.

The Tip List registers (1) the WLN symbol, (2) its valence, and (3) the number of bonds which it has made. The first entry to the Tip List (initial WLN symbol) is assigned its normal valence except for alkyl numerals which are entered with valence of 1. Since the initiating symbol cannot be branched, it must use all of its valence in bonding to the second symbol (*e.g.*, NCR). As each subsequent symbol is encountered, it enters the Tip List using the least number of bonds permitted. For example, see Table III. In the first example, the C entered the Tip List by using 3 of its 4 bonds because all of N's bonds had to be used or else it would have become a branch point. In example 2, the U cannot be a branch point and thus C enters using the least number of its bonds possible, 2. In the third example, Y can be singly bonded and so C uses only 1 bond on entry.

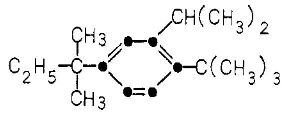
Table III

	WLN	Symbol	Valence	Bonds	Step
(1)	NCR ↑	N	3	3	1
	NCR ↑	C	4	3	2
(2)	1UCU1 ↑	1	1,2	1,2	1
	1UCU1 ↑	U	4	2,4	2
	1UCU1 ↑	C	4	2	3
(3)	Z1YCN&202 ↑	Z	1	1	1
	Z1YCN&202 ↑	1	2	1,2	2
	Z1YCN&202 ↑	Y	3	1,2	3
	Z1YCN&202 ↑	C	4	1	4

As soon as the bonds of a symbol equal its valence, it is cancelled from the Tip List. (An exception is made for a normally nonterminal symbol which is rendered terminal because of charge assignment. It stays until the "&" terminating the branch removes it; *e.g.*, the "O" at position 12 in example I.) Unless they initiate the WLN, the terminal symbols as defined in Rule 8a[10] are not actually entered into the Tip List.

By far the most frequent occurrence of an uncharged S with valence 4 occurs in conjunction with oxygen as S→O or SO_2. Therefore, when S is encountered in the decoding scan, the program examines the adjacent symbols. If the S were immediately preceded by a W or initial O, or immediately followed by a W, a terminal O (*i.e.*, O&), or a ZW, then the S would enter the Tip List with a valence of 4.

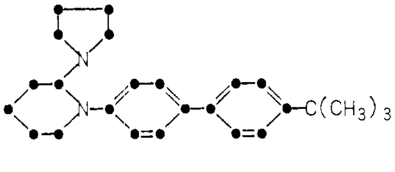
Spaced letters (*i.e.*, locants) specify the point of attachment on a previously cited ring of the substituent which follows, but they also have an important effect on *previous* entries to the Tip List. When a locant is encountered in the decoding scan, it drops all tips back to, but not including, a ring. The program allows for two exceptions where the locant refers to the ring which *follows* it: (1) because of a multiplier, *e.g.*, 2OVR 3 C ER where the "C" and "E" locants refer to the second R; and (2) as the locant for the at-



	WLN	Tip	Valence	Bonds
(1)	2X&&R CY DX ↑	R	6	1,2
		Y	3	1
(2)	2X&&R CY DX	R	6	1,2
	drops →	Y	3	1
(3)	2X&&R CY DX ↑	R	6	1,2,3
		X	4	1

tachment of a ring, *e.g.*, T6NTJ B- AT5NTJ where the locant "A" is in the five-membered ring.

Unspaced ampersands also have a special effect on the Tip List in assigning branch attachments. A group of one or more unspaced ampersands *immediately followed by a spaced letter*[‡] has the effect that all tips back to, but not including, a ring are dropped from the Tip List; and, *in addition,* a tip is dropped for each ampersand in the group. An example is



T6NTJ AR DR DX&& B- AT5NTJ

where the two ampersands and "space B" following the "X" clear the X and the two "R" symbols from the Tip List so that the two "T" rings can be connected.

If the group of one or more unspaced ampersands is *not* followed by a spaced letter, each ampersand is processed separately. Its effect, depending upon the latest entry to the Tip List, is shown in Table IV.

ADJUSTMENTS TO MOLFORMULA TABLE

Methyls. Since methyl-contracting branch symbols (K, X, and Y) increased the molformula count as if they were completely methyl-substituted, each bond to such a symbol in the Tip List causes the reduction of molformula count by C_1H_3.

Hydrogens. A hydrogen is subtracted from the molformula count: (1) if a Y symbol outside a ring is adjacent to a U symbol, *e.g.*, G2UY representing $ClCH_2CH{=}C(CH_3)_2$; (2) for each bond to an alkyl numeral; (3) for each bond to a ring atom except those in L. .J or T. .J rings which are heteroatoms, or Y. The ring dissection routine (see following article) generates, for each ring kernel, a table of all the locants in the ring system together with the atoms occupying each locant position. When ring substituents are encountered, the locant specified for the substituent is checked against this table to determine if an "H" is to be subtracted. In processing any WLN where a locant precedes the ring to which it refers, the locant is saved until the ring has been decoded, and then the "H" is removed if necessary.

MULTIPLIERS

Acyclic multiplied groups are simply rescanned until the group has been entered the appropriate number of times. Admittedly this is inefficient and could be refined. The procedure is currently limited to unnested multipliers. There are four basic types of acyclic multiplied groups.

(1) Initial. The multiplied group initiates the notation and is attached in mirror image order either to itself or to a central symbol. In this case there are no slashes immediately preceding or following the multiplier, *e.g.*, 1VO1 3X2 representing $(CH_3CO_2CH_2)_3CCH_2CH_3$.

(2) Interior. When the repeated group is part of a chain, the multiplier is immediately preceded by a slash; if the repeating group does not initiate the WLN, there will be another earlier slash, *e.g.*, ZSW/2VM/ 43 representing $H_2NSO_2[CH_2CH_2C{=}O(NH{-})]_4CH_2CH_2CH_3$.

(3) Final. The multiplied group does *not* initiate the WLN but is attached to a branch symbol which, except for the multiplied group, would end the WLN. It is characterized by a slash preceding, but not immediately preceding, the multiplier, *e.g.*, Z1U1X/101 3 representing $H_2NCH{=}CHC(CH_2OCH_3)_3$.

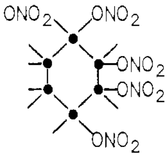
(4) Both Ends. Two or more multiplied groups are attached to an asymmetric central core at the outer symbols of that core (and at any remaining bonds on branched core symbols). This type is characterized by a multiplier immediately *followed* by a slash, followed at a distance by a second slash, *e.g.*, MUYZMNU 2/Y&1/ representing $HN{=}C(NH_2)NHN{=}C(CH_3)CH{=}NNHC{=}NH(NH_2)$. Our present program does not handle multipliers greater than 2 in this category.

Before rescanning the multiplied groups of types 1, 3, or 4, all entries in the Tip List due to that group are dropped. At the conclusion of the final scan, information in the initial or final symbol is retained for the attachment to the unmultiplied segment. Type 2 is handled just as if the notation were uncontracted.

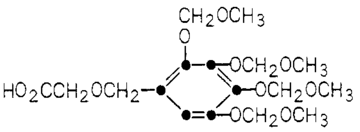
Multiplied ring substituents are also of four basic types.

(1) Identical substitution on *all* available positions, *e.g.*, L6TJ-/G 6 or G 6-R. The program subtracts H_6 and adds Cl_6 to the molformula count.

(2) The multiplied groups' positions must be specified and they do not initiate the notation, *e.g.*

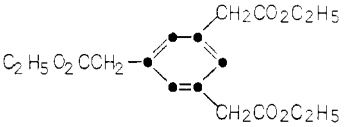


L6TJ A- A- B- C- D-/ONW 5

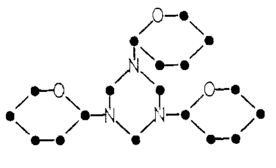


QV101R B- C- D- E-/0101 4

(3) The multiplied groups' positions must be specified and they *do* initiate the notation, *e.g.*



2OV1 3 C ER



T6OTJ B- 3 A C ET6N CN ENTJ

[‡] Spiro ring are designated in a way which must be differentiated from the usual unspaced ampersand class. The sequence "L-&" appearing after a ring system is recognized as a spiro ring junction and treated as a chain of rings with the overlapping atom being entered into the molformula table only once even though it may appear in the notation several times. For example

L6XTJ A-& AL5XTJ

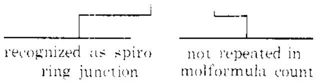
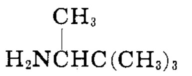
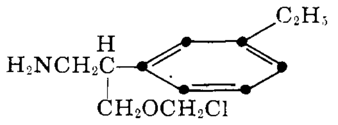
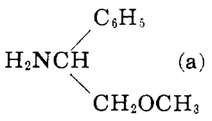
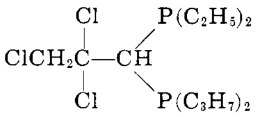
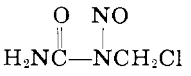

recognized as spiro ring junction
not repeated in molformula count

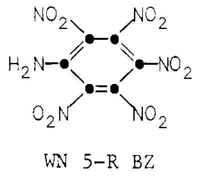
Table IV

Latest (active) entry	Effect of unspaced "&"	WLN	Tip	Valence	Bond	
(1) K, X, or Y	Add a bond. Drop entry if fully bonded.	ZY&X	Z Y	1 3	1 1,2	
(2) Alkyl chain or any fully bonded symbol	Drop it. Drop previous entry also if it is a ring.	Z1YR C2&101G drop →	Z 1 Y R 2	1 2 3 6 2	1 1,2 1,2 1,2 1	
(3) Ring or other variable-valence branch	Drop it.	ZYR&101 drop →	Z Y R	1 3 6	1 1,2 1	 (a)
		G1XGGYP3&3&&P2&2 drop →	. . Y P 3 3	. . 3 5 2 2	. . 1,2 1,2,3 1,2 1,2[a]	 (b)
(4) All others (*e.g.*, the O in terminal NO group)	Make all remaining bonds from the most recent to the previous entry; drop most recent; drop previous, if fully bonded.	ZVNNO&1G drop →	. . N N O	. . 3 3 2	. . 1,2 1,2,3 1,2	

[a] Note that the "&" following each "3" dropped it from the list. The second "&" following the second "3" dropped the "P" which was the latest active entry, having two bonds yet to be used.

The "space-locant-space" which follows the multiplier calls for the program to count such locants appearing before the ring symbol. If their number is less than the multiplier, then the locant "A" is added.

(4) The multiplied groups initiate the notation but all positions are occupied, *e.g.*



WN 5-R BZ

The program processes the five NO_2 groups and adds them to the molformula count, then adds the "R," subtracts H_5, and finally processes and adds the "Z."

LIMITATIONS

The limitations of the present program are these:§

(1) In structures with branching chains of nonbenzene rings (where the branch also contains a nonbenzene ring), only one symbol table is saved and so the hydrogen count may be erroneous.

(2) Nested multiplied groups are not processed.

(3) An unspaced "D" outside a ring is valid only as a D-ring initiator.

(4) For multiplied groups on both ends of an asymmetric central core, the multiplier can only be 2.

(5) When a multiplied ring substituent, cited without a locant, is attached to a heteroatom, a hydrogen is removed.

(6) The sequence "space L- space" is never interpreted as a methyl contraction.

The procedure described has been implemented in PL/I on an IBM 360/40 computer at Pomona College. Using both an "in-house" file of structures and also the "Common Data Base" (National Technical Informational Service, Springfield, Va. 22151) as a test, the program decoded and checked molformulas of acyclic structures at a rate of about 150/min.

ACKNOWLEDGMENT

This work was supported by Grant CA 11110 from the National Institutes of Health Contract NIH DRR-70-4115 from the National Institutes of Health and the Pomona College Medicinal Chemistry Project.

§ It is possible to program the detection of syntax errors which do *not* affect molformula. For example, programming Rule 2 which requires the notation for unsymmetrical structures to begin with the latest alphanumeric symbol is straightforward even when the four types of multipliers are considered. Errors of this type in acyclic notations frequently involve the application of Rules 6 and 7. As soon as the revised form of these rules has been approved by the Notation Association membership, it is planned to implement them in the checker program.

LITERATURE CITED

(1) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968, p 14.

(2) Palmer, G., "Wiswesser Line-Formula Notation," *Chem. Brit.*, **6**, 422–426 (1970).

(3) Campey, L. H., Hyde, E., and Jackson, A., "Interconversion of Chemical Structure Systems," *Chem. Brit.*, **6**, 427–430 (1970).

(4) Granito, C. E., and Garfield, E., "Substructure Search and Correlation in the Management of Chemical Information," *Naturwissenschaften*, **60**, 189–197 (1973).

(5) Feldman, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *J. Chem. Doc.*, **11**, 154–159 (1971).

(6) Miller, G. A., "Encoding and Decoding WLN," *J. Chem. Doc.*, **12**, 60–67 (1972).

(7) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.*, **11**, 52–59 (1971).

(8) Bowman, C. M., Landee, F. A., Lee, N. W., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures," *J. Chem. Doc.*, **8**, 133–138 (1968).

(9) Reference 1, p 237.

(10) Reference 1, p 16.

Computerized Management of Structure–Activity Data. III. Computerized Decoding and Manipulation of Ring Structures Coded in WLN

DAVID ELKINS,[†] A. LEO,[*] and CORWIN HANSCH

Department of Chemistry, Pomona College, Claremont, California 91711

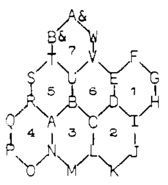
Received December 21, 1973

Construction of the locant path, a table of interatom connections, and the molformula from the WLN for complex ring systems (including polycyclic, spiro, perifused, bridged, and pseudo-bridged rings) is described. The procedure has been programmed for a digital computer and has been found useful in manual decoding also.

WLN Ring Code analysis can be broken down into three principal tasks: (1) dissection of the WLN into its constituent parts (ring locants, ring numerals, nonconsecutive locant pairs, bridges, etc.), (2) construction of the locant path and a table of interatom connections, and (3) molformula calculation. The first step presents no real problem to the experienced human decoder and it turns out to be a straightforward task in programming. For this reason it will not be discussed in this paper.

LOCANT PATH

The guiding principle which underlies the procedure for developing the locant path from the WLN is simple: extend the locant path by using successively later alphabetic symbols, *only after* making sure that the information already processed has not established part of the particular ring pathway being sought. In other words, in the later rings of a fused ring system, it is often possible to get a "running start" on a new ring by using the locants which are common to some of the earlier rings.



1

T D6 C6666 B6 T6 5ABCDU B&J

In the fused ring system 1, the pathway begins at "A" but does not complete its first ring until it reaches "I." The second ring begins with the earliest locant, "C", and continues to later (not always *successively* later) locants, as will be explained below, until the ring contains the number of atoms specified. This also holds true for rings 3 and 4. Ring 5 begins with "A" and the *forward* path is along the route of *latest* previous connection which is to "R." The six-membered ring is not completed with "V," however. The information already in hand that the bonds to "A" were fulfilled with ring 4 should have been used to take a "running start" at ring 5 and begin it with "B." Similarly, ring 6 had "B" as its earliest locant, but the information already processed leads us to begin its running start as E–D–C–B and to complete the ring with "U" and "V."

This procedure is useful for both automatic and manual decoding. It has the advantage over graphical methods in that the structural diagram need not be attempted until the analysis is complete. Granito, *et al.*,[1] mention a similar approach in converting WLN to Ring Codes. Although it is a simple principle, its application is somewhat complex if it is to be applied by computer to the most complex perifused ring systems, as will be seen in the following section.

The first step in the process is the creation of a seven-column "Ringdecod" table containing a row for each locant in the notation. See example number 1. (It will be readily apparent that for manual use this rather formidable table can be greatly simplified or dispensed with altogether. It is a convenient format for computer use and for explanatory purposes only.) The row order is "A" through "W," "A&" through "W&" followed by the cited branch locants in the same order. If the "last locant" is not cited in the WLN, an estimate can be made and space left for possible additions, or it can be calculated from the formula[2] $n = s - 2(r - 1) - b$, where n = the numerical equivalent in the alphabet of the "last locant," s = the sum of the ring numerals in the notation, r = the number of ring numerals, and b is the number of branch locants. The contents of the columns in the "Ringdecod" table are as follows:

[†] Present address: G. D. Searle & Co., Chicago, Ill. 60680.

[*] To whom correspondence should be addressed.