

- (24) Eliel, E. L., "Stereochemistry of Carbon Compounds," pp. 92-4, McGraw-Hill, New York, 1962.
- (25) Hendrickson, J. B., Cram, D. J., and Hammond, G. S., "Organic Chemistry," 3rd ed., pp. 204-6, McGraw-Hill, New York, 1970.
- (26) March, J., "Advanced Organic Chemistry: Reactions, Mechanisms, and Structure," p. 84, McGraw-Hill, New York, 1968.
- (27) Morrison, R. T., and Boyd, R. N., "Organic Chemistry," 3rd ed., pp. 130-33, Allyn and Bacon, Boston, 1973.
- (28) Noller, C. R., "Chemistry of Organic Compounds," 3rd ed., pp. 368-70, Saunders, Philadelphia, 1965.
- (29) Hendrickson, J. B., "A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis," *J. Amer. Chem. Soc.* **93**, 6847-54 (1971).
- (30) Davis, C. H., "A Simple Code for Improving the Retrieval of Information Associated with Keto-Enol Tautomers," *J. Chem. Doc.* **6**, 199-205 (1966).

A Qualitative Comparison of Wiswesser Line Notation with Ringdoc

MITSUO SASAMOTO, TAKASHI KUBOTA, TOSHIKI HAMANO, TAKESHI SHINBA, and MASAKAZU NAKAI
Information Center, Tanabe Seiyaku Co., Ltd. 2-2-50 Kawagishi Todashi, Saitama, Japan

Received September 7, 1973

Two systems, WLN and Ringcode, for retrieving structural information were analyzed qualitatively and evaluated for a series of chemical compounds. The studies ranged from specific to generic questions and also involved retrieval by fragments. Neither system was completely satisfactory for all types of searches.

In the field of chemical and pharmaceutical sciences, a large part of the literature is related to chemical compounds. In fact, some 85% of the index entries in the 1966 Subject Index to *Chemical Abstracts* were associated with compounds and materials, according to the CAS survey. It is estimated that the total number of known chemical substances is some four to six million, and additions are appearing at the rate of some 150,000 to 250,000 per year.

In chemical information management, special codes or notations are used to store and retrieve compounds. Methods of representing chemical compounds fall historically into two groups: conventional and nonconventional. The former are based mainly on nomenclature, such as chemical names, trivial names, proprietary names, and trade names; these are the conventional indexing terms used by chemists. Chemical structure diagrams and molecular formulas also can be included in the conventional group. Though word-based methods are suitable for representing chemical compounds both in printed media and oral communication, they are quite inconvenient and almost useless as a general tool for substructure searches—those concerned with partial rather than exact matching of descriptions.

Nonconventional methods of representing chemical compounds usually fall into three classes: fragmentation codes, linear notations, and topological codes. All three have been developed with the mechanical aids of the 20th century—PCS (Punched Card Systems) in the 1950's to early 1960's, and EDPS (Electronic Data Processing Systems) afterwards. In these nonconventional systems, each compound is considered as a composite of fragments—rings, functional groups, connections of atoms, or the like.

In Japan, the organizing of chemical information systems with electronic data processing (EDP) equipment received great impetus in most pharmaceutical firms through the introduction of the Ringdoc system offered by Derwent Publishing Co. London, in 1964. There are two types of magnetic tapes available in Ringdoc: term search tapes (Codeless Scanning) and fragmentation-code tapes (Ringcodes). The latter is suitable for chemical structure searches, with the realization that Ringdoc is a closed system, rigidly frozen by the fragment definitions. Seventeen pharmaceutical firms, including Tanabe Seiyaku Co. in Japan, presently subscribe to the Ringdoc tapes.

Tanabe Seiyaku Co. subscribed to this service in 1965

and set up an information retrieval system that included the company's internal file of compounds. In 1969, we also started to encode these internal compounds by Wiswesser Line Notation (WLN). By the end of 1972, some 20,000 new compounds were registered in both files, Ringcode and WLN. In this paper, a qualitative analysis and an evaluation of Ringcode and WLN searches on their respective internal files are described. Some WLN problems also are discussed.

CHEMICAL STRUCTURE REPRESENTATION BY RINGCODE AND WLN

Ringcode is a fragmentation coding system in which three different types of structure descriptions are available—general, steroid, and peptide. Hereafter, the term Ringcode will be used to mean the *general* code. Each compound is fragmented according to predetermined concepts, such as type or size of rings, kind or number of heteroatoms, kind of functional groups, length or type of carbon chains, and the like. These fragment distinctions are punched into tab cards in a binary mode, that is, by a one-hole, one-meaning method. Columns 2 to 27 are used on the tab card for the general chemical code; column 1 is used for identification codes that discriminate the kind of codes to be used thereafter.

Fragmentation codes such as Ringcode usually are not unique and unambiguous by nature, because the records do not provide places to show how the fragments are connected, and the chemical structure cannot be regenerated from the fragments without this essential assembling information. However, one of the merits of the Ringcode is that rather fast search times can be achieved by assigning a corresponding bit code for each fragment in the computer records, using this binary mode (*zero* meaning absent and *one* meaning present). This maximum efficiency with binary searching is, of course, applicable to any bit screen like those in Ringdoc, including screens generated by computer processing of the symbols in the WLN records.

WLN, like traditional line-notations, delineates chemical structures exactly as the fragments are connected, and it cites the connecting positions on rings. Unlike traditional line formulas, only *one* citing order is allowed, so that the description is **unique** as well as unambiguous. The symbol set, and the basic citing rules for this nota-

tion were reported by W. J. Wiswesser in the early 1950's. The rules were greatly expanded and generalized by E. G. Smith¹ and other collaborators in a 1968 revision. WLN now is the most widely used linear notation in the world: It is used in external data base services such as the *Index Chemicus* Registry System (ICRS) tapes of *Current Abstracts of Chemistry and Index Chemicus* (CACIC), produced by ISI (Institute for Scientific Information), and the Drugdoc service of Excerpta Medica, as well as many internal data bases.

Every detail of structure is covered by the 40 WLN symbols that are all available in standard computer and tabulating equipment: upper case letters, numerals, and just four punctuation marks (ampersand, hyphen, slash or virgule, and asterisk). The most used signal is the space, which functions as a "shift key" to provide lower-case meaning of letters (to locate ring positions), at the same time setting off these logical units of information that begin with a locant. Unspaced letters generally denote atomic groups in the traditional manner, and unspaced numerals denote chain or ring sizes.

In WLN encoding, there are strict rules that determine the sequence of symbols, aimed to preserve uniqueness of the citing path and to obtain optimum use of the alphabetic symbols. This notation is suitable for input to computer systems, and is particularly useful for several kinds of computer-generated indexes, such as ring derivatives and compacted listings (subdividing the alphabetic sequences by notation size or length) display hundreds of descriptions on a single sheet of line-printed paper.

String searching of any kind of description requires much more computer time than bit searching, because the bit signals are registered in the computer's basic binary language. Thus string searching of WLN descriptions is far less efficient (by a factor of *thirty to fifty*) than binary searching, like that in Ringcode, done on bit screens generated by the computer program from the WLN symbol-scanning. The over-all gain in using WLN bit screens was reported by Granito *et al.* in 1971.²

SEARCH LOGIC IN COMPUTER PROGRAM

Search results are considerably influenced by the design efficiency of the computer program—and especially by the search logic, so the types of search logic available in our programs must be explained before discussing comparative results.

Our Ringcode programs use the following types of logic:

Card Type	Type of Logic	Card Ident. Punch (Col/Row)
1	AND	80/1
2	OR (including simple OR and major OR)	80/4-80/8
3	NOT (absolute)	80/2
4	NOT (conditioned)	80/3

The rest of a "Type 1" card typically has punches for all of the fragments (bits of structure) that are *wanted* in the search question; in contrast, a "Type 3" card has punches for the bit details that are *not wanted* under any condition, such that all records containing *any one* of these unwanted details are ignored. On a conditional "Type 4" card, a plurality of unwanted details means that the corresponding records are ignored only when they contain *all* of these "Type 4" unwanted bits. Our experience shows that the careless use of *NOT* logic in fragmentation code (Ringcode) searching involves the risk of missing relevant information.

Our WLN programs employ the following types of logic on terms that typically are strings of WLN symbols:

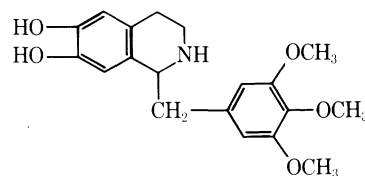
Type of Logic	Comment
Simple AND NOT	Each "AND" term is assigned a different term weight, including absolute NOT and conditioned NOT. Both function by an assigned term weight that prevents any "overwhelming" (achievement of the threshold weight) by other terms.
Simple OR	Equal term weight is assigned to each term.
Major OR	Different "OR" groups are separated by MAJOR "OR," with equal term weight for those terms in the same group.
Followed By (FB)	The assigned term is required to follow another in a sequence of terms.
Same Word (SW)	"SW" is used when two or more terms must be within a symbol sequence that has no separating space(s).
Any Locant (*)	The asterisk (*) mark means a letter following a space—that is, a ring locant in WLN descriptions.

EVALUATION OF WLN AND RINGCODE RETRIEVALS

Specific Search or Exact match Specific searches often are done when a chemist wishes to confirm that his newly synthesized compound is indeed new and not in the file. In all such cases, a quick answer is obtained from a computer-generated and alphabetized WLN list, or from a permuted WLN index.

Figure 1 shows the comparative results of a specific computer search. While the exact match of the WLN for 3,4,5-trimethoxybenzene derivatives gives no false hints from WLN, such false matches or noise occur in Ringcode (for example, the 2,3,4-derivatives). However, for moderately complex compounds, our searching results showed about the same relevance ratios for Ringcode as for WLN, because the complexity provides more details (more specificity) in specific searches.

Substructure Searches for Partial Matching. Most computer-managed chemical searches are related to substructure questions, seeking all compounds that contain the specified substructure. The capabilities of such searching at reasonable costs and speeds by simple procedures are important factors in the handling of chemical



WLN: T66 CMT&J BIR Cöl Döl Eöl& HQiQ	Relevance 100%
Ring code: 2/0: total number of rings in the molecular: 3	91%
(main punch positions only) 3/1, 6/1, 7/1, 8/1:	
9/4, 9/9: interrelative positions of substituents on benzene rings	
10/2: position of substituent on hetero ring	
11/5: one —CH ₂ —	
18/1: two —OH	
18/12, 18/3: two or more —OCH ₃	

Figure 1. Specific search

(a) $-\text{CO}-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{CH}_2-\text{CO}-$	Relevance
WLN: V4V or V024	100%
Ringcode: 11/1, 11/8: four $-\text{CH}_2-$ (11/8 includes 4 \sim b $-\text{CH}_2-$)	15%
23/1, 23/7: $-\text{CO}-$	
(b) $\text{C}-\text{CH}-\text{CH}_2-\text{NH}-$ $\begin{array}{c} \\ \text{O} \\ \end{array}$	
WLN: Y δ (FB)IM(δ R) δ Y(FB)IM(δ R)MIY(FB) δ	90%
Ringcode: 11/5: one $-\text{CH}_2-$, 13/1: X $-\text{C}-\text{C}-\text{X}'$	70%
18/4: $-\text{O}-$ 12/6: $>\text{CHX}$	
19/2: $-\text{NH}-$	

Figure 2. Substructure search—1

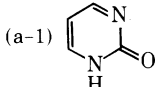
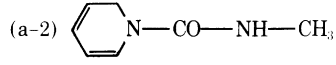
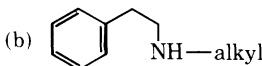
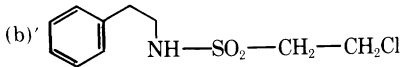
(a) $-\text{NH}-\text{C}-\text{N}-$ $\begin{array}{c} \diagup \\ \text{O} \\ \diagdown \end{array}$	Relevance
WLN: MVN(δ R)N(SW)VM	43%
Ringcode: 24/3: TN $-\text{C}-\text{NT}'$	80%
24/7: Y= δ	
26/12: T=H	
26/11: T'=R	
(a-1) 	
T6MVNJ	
(a-2) 	
T6N BHJ AVMI	
(b) 	
WLN: R(FB)2M(δ R)MZR	30%
Ringcode: 3/12, 3/2: one isolated benzene	60%
11/11, 11/6: two $-\text{CH}_2-$	
19/2: $-\text{NH}-$ (amine)	
(b') 	
G 2 SWM2R	

Figure 3. Substructure search—2

information, yet there are no obvious and ideal traditional solutions to this general problem.

Figures 2 and 3, for example, show that the relevance ratio of both Ringcode and WLN varies with the type of question. Thus, an over-all evaluation cannot be made; the evaluation here is limited to the questions asked, and should be extrapolated to compare Ringcode and WLN in general. Thus, for the $-\text{CO}(\text{CH}_2)_4\text{CO}-$ fragment, WLN gives a sharply defined V4V string description [wherein V means carbonyl and 4 means the *unmodified* $(\text{CH}_2)_4$ chain]. In the Ringcode search, however, the relevance ratio was as low as 15%, because in this system the two fragments (carbonyl group and the fixed definition of 4 to 6 $-\text{CH}_2-$ groups linked together) occur with high frequency, yet must be searched independently.

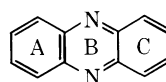
Figure 2b is a type of question involving the common $-\text{CH}-$ branching group. Since WLN may start at the end of any one of these branches, all possible permutations of sequences must be tried—six for this case. Therefore, careful planning of search strategies is needed, especially for searches containing branches of any kind—the ternary and quaternary carbon-branches, the corresponding nitrogen-branches, and even the branched benzene ring, since this is treated as a branched-R symbol in WLN. The

sharpness of a WLN string sequence is sensitive to direction, whereas the bits are not.

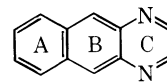
Figures 3a and 3b are examples showing better results for Ringcode than for WLN. In the Ringcode, the punch positions for each fragment of $-\text{NH}-\text{CO}-\text{N}-$ or the corresponding thiourea were suitably assigned. In the WLN search, two alternative terms had to be connected with OR logic to cover both citing directions of the symbol string: $\dots \text{MVN} \dots$ and $\dots \text{N} \dots \text{VM} \dots$ sequences. The second case is expressed as N(SW)VM to show that the parts are in the "same word," and are not separated by a space. In this open chain search, pyrimidones (Figure 3a-1) were retrieved because the $\dots \text{MVN} \dots$ or $\dots \text{NVM} \dots$ sequence is present in them; hence, these were unwanted, and truly branched-N compounds like that in figure 3a-2 were missed by this statement of the search question, because here the parts are separated by spaces (for the ring locants).

SEARCHES INVOLVING SPECIFIED RING POSITIONS

Cyclic positional specifications can be WLN advantages as well as the above-noted disadvantages. Often the position and interrelationship of *substituents on rings* and *heteroatoms in rings* are important in substructure searches. Ringcode generally is unsatisfactory in showing such relations because its fragmenting approach is not sufficiently specific; it assigns some 12-punch positions for positional relations among heteroatoms (namely, the 12-row in column 8) and 24-punch positions for the same among substituents (both 12-punches in columns 9 and 10). These fragmentary descriptions obviously cannot indicate specific relationships—e.g., between heteroatoms and substituents—when there are many heteroatoms or substituents; this inevitable uncertainty occasionally generates undesirable answers from Ringcode. For example, in a search for PHENAZINES, the answers will include BENZOQUINOXALINES, because both contain the *para*- or 1,4-separation of heteroatoms, and in Ringcode no complete discrimination can be made on the multitudes of positional relations in polycyclic systems, such as the 1,4-separation in the **center** ring of a tricyclic system *vs.* that same separation in an **end** ring of a tricyclic system.



Phenazine



Benzoquinoxaline

WLN descriptions are designed to locate substituents on rings and heteroatoms in rings unambiguously, and in a strikingly simple mechanical manner: a *space* is used as a shift key (a symbolic *operator*) to convey **lower-case meaning** to the letter that follows, and these virtual **LOwer CAse letTErs LOCATE** the heteroatoms in ring systems and (after the ring-closing J-mark) the substituents on them. These spaces also set off the **units** of cyclic information in a logical flow, with the locants of substituents appearing between the ring symbols and the substituent symbols.

Figure 4a illustrates a search for N-substituted INDOLES. Here the WLN locant chain starts at one of the fusion points (so both rings have this lowest possible position), and specifically at the one next to the N-atom (so it has a lowest possible position in the ring-tracing path, relative to that starting point). Thus the WLN string T56 BNJ B \dots sharply specifies simple N-substituted INDOLES, and $\dots \text{BT56 BNJ}$ shows this fragment connected to a larger or higher-ranking ring at that same B-position. (The first string also covers N-substituted INDOLES similarly connected at any other position beyond the B-position.) We found that our search strategy for

A QUALITATIVE COMPARISON OF WISWESSER LINE NOTATION WITH RINGDOC

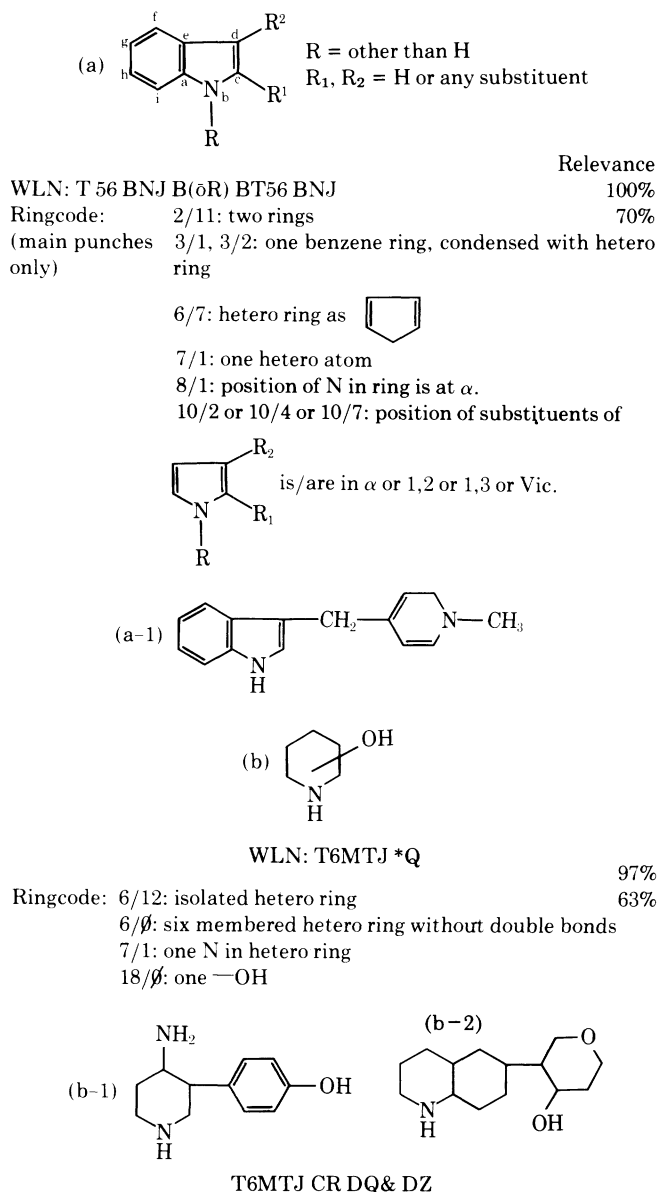


Figure 4. Location searches on the substituents on ring

WLN strings was unsatisfactory as phrased for notations involving multipliers, but in practice the number thus missed was negligible. Alternatively, WLN coding could be done without the multipliers, or the questions could be phrased to handle multipliers as new alternatives. Figure 4a-1 shows an example of typical noise by Ringcode, here failing to state substitution on the nonaromatic NH-group.

Figure 4b shows a search on x-HYDROXYPIPERIDINES, with the generic mark (asterisk) used in the WLN search logic to mean ANY POSITION. In this rather simple case, the locant must be just 2, 3, or 4, so in Ringcode the corresponding punch positions (10/3, 10/4, or 10/5) were coded as logical alternatives. Figure 4b-1 shows the type of structure that represented noise both in Ringcode and WLN, while figure 4b-2 shows the type that represented noise only in Ringcode, with less specific interrelations.

SEARCHES BASED ON RING SKELETONS

There are some 30,000 kinds of ring skeletons, so it is mathematically impossible to distinguish closely related types with just some hundred fragmentary specifications.

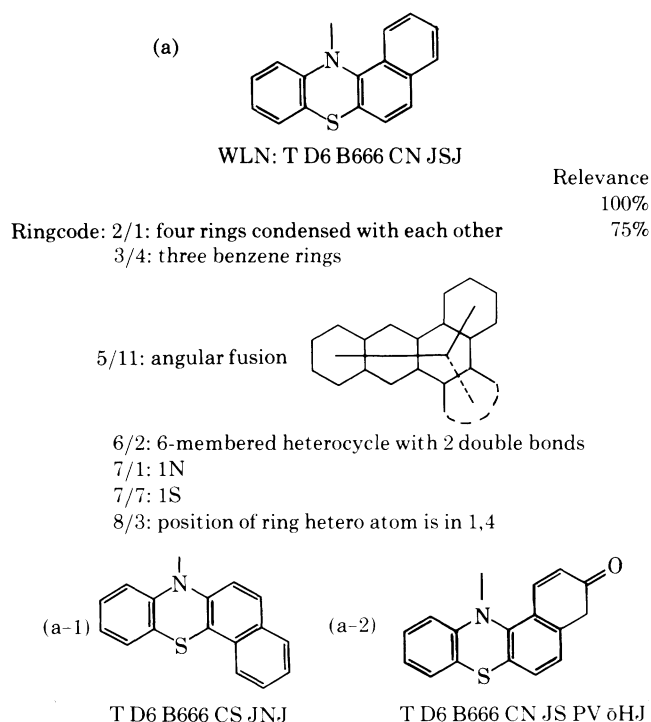


Figure 5. Search based on ring skeleton

Figure 5 shows that the illustrated BENZOPHENOTHIAZINES are sharply defined by the WLN string, T D6 B666 CN JSJ, and no different type of skeleton can match this search definition. In the less specific Ringcode, it is difficult to exclude closely related isomers like that in figure 5a-1 (with reversed heteroatoms, as shown in the WLN descriptions). However, if cyclic keto-groups are accepted as relevant answers to the question, the original Ringcode specification still applies while the WLN searching specification must be modified to allow this new variable as an answer.

GENERIC OR MARKUSH-TYPE SEARCHES

Ringcode specifications are by nature fragmentary or, in a way, generic, so it seemed to us that Ringcode was more suitable for Markush searches than WLN. Figure 6 illustrates the superiority of bit screen searching over string searching when alternate substituents cause changes in the ordered direction of the string citation. In WLN, the starting point is the variable when it is NH₂ (high-ranking Z symbol) or NO₂ (with W for the O₂), but not when it is —CH₃ of the methoxy group. Of course, the additional faster and cheaper searching of bits gives these a fundamental advantage over the searching of strings; and in such generic searching, Ringcode has no inherent advantage over bit screens of WLN symbols generated by the computer **from the WLN records** when these enter the data base. Figure 7 shows the string-searching difficulties when these three different compounds are to be included in a Markush expression. Another point in favor of processing bits defined by the WLN symbols is that only such bit or fragment searching is workable if a specific compound is sought in the Markush expressions—for example, description (2) from (1) in Figure 7.

PEPTIDE CODING

Polypeptides or amino acid polymers occupy a biologically important and physically unique field in chemistry, so special codes have been devised for their descriptions.

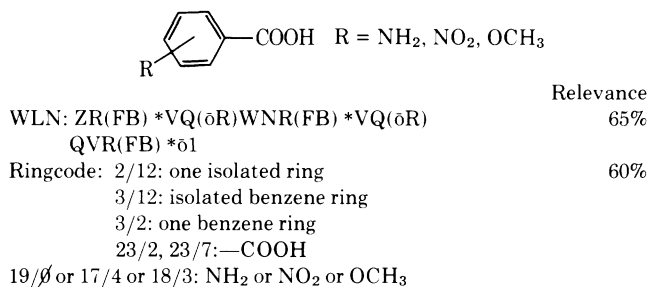


Figure 6. Markush structure search

Full notation: QVR X-A- // -A-/Z/δ1/NW (1)
individual notation:

R = NH₂: ZR XVQ (2)
R = NO₂: WNR XVQ (3)
R = δ1: QVR Xδ1 (4)

Figure 7. Markush-type structure in WLN

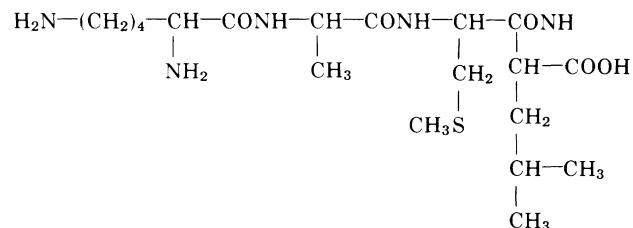
The IUPAC nomenclature, which defines each peptide as a three-letter symbol, has been widely used. For computer-processing methods, the following additional codes or notations are available: the peptide code by Ringdoc, the IC-IUPAC single-letter code of ISI, and the Hoffman modification of WLN.

Smith's manual on WLN contains no rule on peptide coding, but in 1964 Wiswesser was one of the first to acknowledge that "A Protein-Spelling Alphabet," with a single letter for each amino acid residue, is needed and easily designed.⁶ Others subsequently provided a single-letter symbol set officially recommended by the IUPAC-IUB Commission on Biological Nomenclature, and this has been utilized by ISI as described below.

Peptide Code of Ringdoc. One column of an IBM card is assigned to each of the 20 amino acids, arranged alphabetically by name—alanine in column 2, arginine in column 3, aspartic acid in column 4, and so on to column 22. Other amino acids, such as α-aminobutyric acid, lanthionine, norleucine, and the like, are encoded collectively in column 23. Column 24 provides for foreign structures. The digital (0 to 9) punch positions of these columns show the sequence positions of the corresponding amino acids in the peptide chain. Other information, such as the number of amino acid residues in the polypeptide, the types of substituents, types of connections, and ring size, is defined in the remaining columns and punch positions.

Numbering usually starts with the NH₂-terminated amino acid, and this sequence number of each peptide chain unit is punched into the column of the corresponding (alphabetically arranged) amino acid. When more than nine residues are present, the punched digit represents only the **units** value of the sequence number; thus the 7th, 17th, 27th, 37th, etc., positions all are represented by the 7-punch. There is no way to identify fully the sequence relations in this fragment-type code, but the terminal digit method of punching the units values has high discriminating power.

IC-IUPAC/Wiswesser Notation. In this notation,⁴ which was devised for ICRS usage at ISI in Philadelphia, peptide compounds having three or fewer amino acids are encoded in WLN by the standard rules, and the larger peptides (having four or more amino acid residues) are encoded by the IC-IUPAC/Wiswesser notation. The 26 letters of the English alphabet are used to represent the 26 most common amino acid residues, in accordance with the IUPAC-IUB recommendation of 1968, and other resi-



- (1) H-Lys-Ala-Met-Leu-OH IUPAC Nomenclature
(2) 13/1 (Lys), 3/2 (Ala), Peptide code by Ringcoding
14/3 (Met), 12/4 (Leu) IC-IUPAC Code
(3) H///KAML///Q Hoffmann/Wiswesser Notation
(4) Z4YZ/VMY/3 1 1S1 VQ1Y

Figure 8. Peptide coding

dues can be denoted by adding a special character (such as the ampersand) to single letters. Terminal protecting groups are encoded by the standard WLN rules, set off by three slashes from the IC-IUPAC notation, and substitutions within the chain are denoted by two-digit codes (see Figure 8-3). Sequential relations thus are denoted in full, and this notation gives the added advantage of ease in computer searching or in visual scanning.

Hoffmann/Wiswesser Notation. For large peptides, the standard WLN descriptions, as mentioned above, are often very long and tedious. Hoffmann⁵ devised a modified WLN by using multipliers in the notation. In peptide chains, the —CO.NH—CH— linkage appears repeatedly and is denoted as VMY (V means the keto group, M means NH, and Y means tertiary carbon) in WLN, or the reversed YMV. Hoffmann denotes these repeated groups with a multiplier numeral; thus, the three —CO.NH—CH— groups in the peptide shown in Figure 8 is represented as /VMY/ 3. Side chains are encoded by the standard WLN rules, and their citing order (after the multiplier) indicates the connecting sequence of these residues in the peptide chain.

This notation is suitable for peptide searches that involve specific substructures, but in general, visual deciphering is almost impossible. At present, we use the Hoffmann/Wiswesser notation with slight modifications, in registering peptides synthesized in our company, because the descriptions are based on WLN in principle. However, we also utilize printed lists of peptides encoded by the conventional 3-letter IUPAC notation to cover the deficiency of the Hoffmann/Wiswesser descriptions.

CONCLUSION

No single notation or code gives satisfactory answers to those who wish to search completely on chemical structures. For establishing our internal chemical registry system, we at first adopted the Ringcode and then WLN. In the next development, we are going to combine these records—with close attention to cost reductions possible with WLN bit screens²—to improve the efficiency of searching methods and the quality of the results while expanding the existing system.

ACKNOWLEDGMENT

We wish to thank Wm. J. Wiswesser and Charles E. Granito for their helpful suggestions during the preparation of this paper.

We are indebted to H. Miyagishima and N. Hoshino for their assistance in the experimental work.

LITERATURE CITED

- (1) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N.Y., 1968.
- (2) 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(2), 106-10 (1971).
- (3) Ringdoc Instruction Bull. No. 7, "Peptides," 1964 (for subscribers only).
- (4) Revesz, G. S., "One-Letter Notation for Calculating Molecular Formulas and Searching Long-Chain Peptides in the *Index Chemicus* Registry System," *J. Chem. Doc.* 10(3), 212-15 (1970).
- (5) Hoffmann, E., "Use of a Modified Wiswesser Notation for the Encoding of Proteins," *Ibid.*, 9(3), 137-40 (1969).
- (6) Wiswesser, W. J., "A Protein-Spelling Alphabet," *Chem. Eng. News* 42, 4 (Sept. 14, 1964).

Cambridge Crystallographic Data Centre. IV. Preparation of "Interatomic Distances 1960-65"

F. H. ALLEN, N. W. ISAACS, OLGA KENNARD,* W. D. S. MOTHERWELL, R. C. PETTERSEN, W. G. TOWN, and D. G. WATSON**

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

Received August 20, 1973

The Cambridge Crystallographic Data Centre is concerned with the retrieval, evaluation, synthesis, and dissemination of structural data obtained by diffraction methods. Earlier papers in this series have described the organization of computer-based files of both bibliographic information and of evaluated numeric structural data. The present paper describes the use of a subset of the system to produce a compendium of interatomic distances including stereo diagrams of structures, numeric data, editorial comments, and a variety of indexes.

Earlier papers in this series have described the organization and information content of the bibliographic file¹ and the structural data file² of the Cambridge Crystallographic Data Centre. The process of evaluation of numeric data applied to the latter file has also been documented.³ The bibliographic information is disseminated via the Bibliographic Volumes of the series "Molecular Structures and Dimensions"⁴ (hereinafter called MSD). The computer-typesetting and other techniques used in the preparation of these volumes from our computer files will be described in Part V of this series.⁵

The present paper describes the use of both the bibliographic and structural data files in the preparation of Volume A1 of MSD: "Interatomic Distances 1960-65" (ID).⁶ This volume is a continuation of "Tables of Interatomic Distances and Configurations in Molecules and Ions,"⁷ which covered the literature until the end of 1959. It gave information on interatomic distances and, where relevant, bond angles, obtained by spectroscopy, electron, neutron, and X-ray diffraction for organic and inorganic molecules and ions, and for metals.

In our compilation the criteria for inclusion in the files^{1,2} necessarily restricts coverage to organic and organometallic compounds investigated by X-ray and neutron diffraction, while the spectacular increase in the number of studies since the early 1960's has forced us to consider the period 1960-65 in the first instance. Computer methods have been used both for the evaluation^{2,3} and presentation of results. In most cases, this has enabled us to extend the information given in the original publication by preparing stereoscopic illustrations of individual molecular structures and, where appropriate, by including torsion angles computed from the original data.

DATA BASE

The data base used in the preparation of ID was a merged subset of both bibliographic and structural data files for the chosen period. The file was in the form of card images on magnetic tape. Entries were grouped in 86 chemical classes¹ and ordered by C,H content within each class, as in the bibliographic volumes of MSD.^{4,5} Entries were numbered sequentially in each class⁴ and a card record ENTNO was added to each data entry in the form *m.n*, where *m* is the class number and *n* the serial number. In contrast to the bibliographic volumes where cross-references were provided to supplementary classes, each compound appears once only in ID, in its basic class. Finally a card type IDREF was added to the file, containing bibliographic references to studies of a given compound published outside the period 1960-65. This was generated automatically from the main bibliographic file current at that time.

ORGANIZATION

Responsibility for the data base was divided among six editors who ensured that all entries were complete, undertook any residual scientific checking, and added any additional textual material on the card types REMARK, DISORD, and ERROR² at their discretion. At a later stage, the editors chose a suitable view of the molecule for the crystallographic stereo pair, and edited the tables of numeric data. The responsibilities were allotted on the basis of chemical class, so that each editor had a relatively narrow range of compounds with which to become familiar, e.g., benzenoid compounds, natural products, etc. Frequent discussions were held to ensure editorial consistency.

* External Staff, Medical Research Council.

** To whom all enquiries should be addressed.