On-Line Substructure Searching Utilizing Wiswesser Line Notations †

ALBERT V. TOMEA* and PETER F. SORTER
Hoffmann-La Roche Inc., Nutley, New Jersey 07110

Received July 6, 1976

The scope of the Hoffmann-La Roche integrated interactive chemical information system has been extended by a suite of computer programs written in Fortran for the Honeywell 6080. The programs accept Wiswesser Line Notations (WLN's) and expand those containing contractions into noncanonical notations. In either an on-line or batch mode, the notations are converted into atom connectivity matrices, using "dot-plot" symbols and searched via a series of set reductions. The matrices are generated and utilized for atom-by-atom searching only when bit and symbol string searching of WLN's proves inadequate. Statistics concerning rate of matrix generation and search and the factors influencing these rates are included.

For over ten years, Hoffmann-La Roche has used Wiswesser Line Notations[1] (WLN's) for specific compound identification and for generic structure searches of its confidential file of compounds. The use of rotated indices[2] was quite adequate for these purposes for several years. However, some time ago, a number of factors were identified which convinced us that it was time to develop additional search techniques. These factors were:

(1) Increasing file size and growth rate.

(2) Increasing search requests which were tedious or impossible to carry out using rotated lists of WLN's.

(3) Increased accessibility and decreased cost of on-line computer processing.

(4) Continued availability of about 15 000 WLN's per month from ICRS (Index Chemicus Registry System).

The most important objective was to develop techniques which would make it economically feasible to do atom-by-atom searching. Therefore, a system was designed which stores only the compact WLN. These are subjected to a series of rapid and efficient screening techniques,[3] and only those which filter through are converted to matrices and searched. Upon completion of the search, the matrices are purged.

Notations containing methyl contractions and multipliers do not contain sufficiently explicit information to create atom-connectivity matrices. Therefore, programs were developed which would accept notations containing these space-saving features and expand them so as to provide the necessary information. Although the generated notations may be noncanonical, the nature of the matrices required is such that only semantically correct notations are necessary. This somewhat simplified the problem.

Using Figure 1 as an example, it will be shown that it is possible to (1) develop sufficient explicit information for matrix generation and (2) utilize a semantically correct WLN for matrix generation.

Although structurally there is a plane of symmmetry in the molecule, the WLN rules are such that the locant assignments in the two rings are different. For this reason, tables of equivalent locants are developed based on the ring exit locant of the initial ring (Figure 2). The tables are then stored in the computer for use as needed.

Using the table of equivalent locants for exit locant B, the WLN is expanded as shown in Figure 3. Although the expanded WLN is noncanonical in order, it fully describes a semantically correct WLN (Figure 4).

The only difference between the two notations is the order of citation of the last three substituents. Construction of an atom-connectivity matrix can be accomplished from either.

† Presented at the 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 24–29, 1975.

* To whom correspondence should be addressed.

The second WLN in Figure 4 would be the correct WLN if contractions were eliminated. Recently the Chemical Notation Association voted to remove contractions from the Wiswesser rules.[4] However, since historical files may remain intact, the expansion portion of the program will be retained.

The matrix generator program analyzes the notation and assigns "dot-plot" symbols (DPS)[5-7] to all atoms. Dot-plot symbols impart a unique character based on their bonding patterns (Figure 5).

The asterisk is employed to identify any two-symbol atom cited in the WLN. Discrimination of such atoms is accomplished by the use of molecular formula screens and string searching, prior to matrix generation.

To further characterize each atom, a numerical value is assigned at the time the matrix is being generated. Assignment is based on whether the atom is acyclic, benzenoid, or cyclic nonbenzenoid (Figure 6). The term acyclic is self-explanatory. Benzenoid is defined as a phenyl ring not fused to any other ring. Nonbenzenoid means any other ring including fused phenyl rings. The utility of the cyclic qualifier will be discussed along with search methods.

Analysis of a WLN is accomplished by dividing the notation into two groups: (1) symbol strings which describe cyclic nonbenzenoid structures and (2) all other symbol strings.

For example, each segment in Figure 7 is analyzed as a separate entity. Segment A first undergoes a ring analysis to prepare it for entry into the matrix.[9] Each datum in the segment is analyzed by the ring analysis program to determine its role in the structure description. The symbol string C565 signifies a five-, six-, and five-membered ring with C, A, and A, respectively, as the lowest locant. The two A locants are implied as provided for in WLN Rule 31(b). The program generates locants appropriate to each ring. Each ring is considered independent at this point, although some locants may be common to more than one ring. Assignment of DPS and cyclic qualifiers follows. This crude matrix is sorted into ascending alphabetic sequence, retaining all neighbor connections. Any cited atoms such as those at locants B and F are then incorporated into the matrix.

Although, according to WLN Rule 34, the symbol string SW is cited within ring signs, W is assigned an exocyclic position and acyclic qualifier in the matrix. Positioning the W in this way enables all cyclic locants to retain their original position. For example, by citing the symbol W directly after S in the matrix, locant F would appear improperly positioned in the matrix. All cited endocyclic symbols overwrite the original DPS in the matrix.

Segment B is an acyclic symbol string and subjected to the matrix generation algorithm directly. However, syntactical analysis is necessary to ensure that locants F and H are associated with segment A rather than segment C. This type of analysis is needed when a notation contains more than one

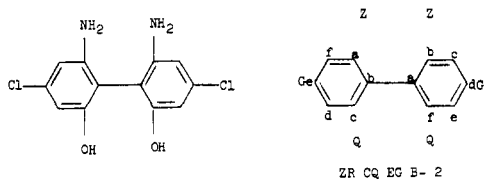


Figure 1.

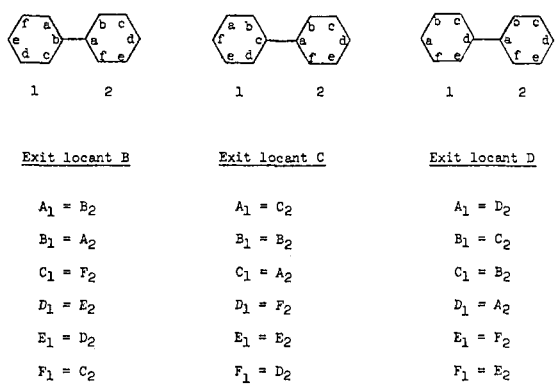


Figure 2.

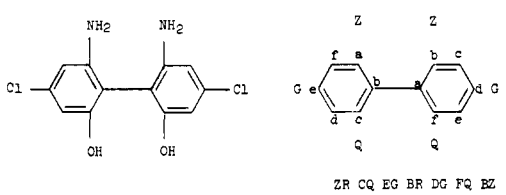


Figure 3.

Non-canonical (Expanded) ZR CQ EG BR DG FQ BZ

Canonical (Expanded) ZR CQ EG BR BZ DG FQ

Figure 4.

cyclic system (including benzene) or symbol chains containing more than one branch symbol. Cited symbols are added to the matrix sequentially. Each substituent is attached to its proper ring locant. Any DPS changes, appropriate connections, and assignment of cyclic qualifiers are made at this time.

Finally, segment C first undergoes ring analysis followed by matrix generation. This matrix is confined to only the ring atoms described by segment C. The proper connections are determined and made along with any appropriate DPS changes. The complete matrix is shown in Figure 8.

Generation of an atom connectivity matrix from a WLN, in effect, creates a two-dimensional model suitable for substructure searching. Using a technique known as set reduction,[8] a program was written to perform atom-by-atom substructure searches which can be either generic or specific. The demands one can make of the system and the search

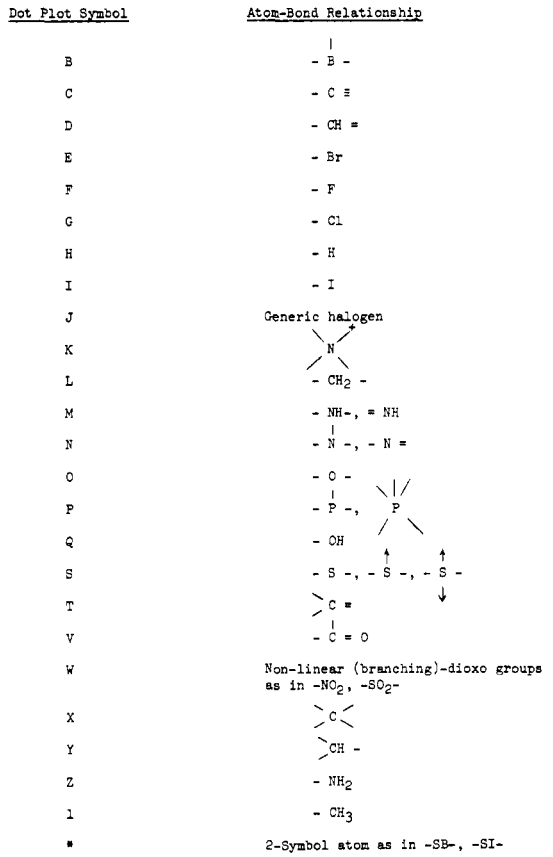


Figure 5.

Cyclic Classification	Qualifier
Acyclic	0
Benzenoid ring	1
Non-benzenoid ring	2

Figure 6.

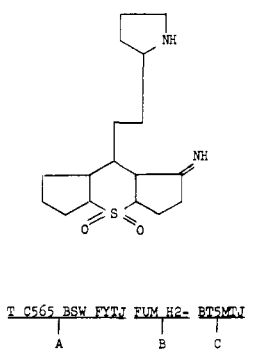


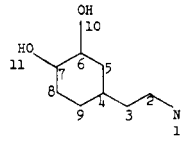
Figure 7.

techniques employed can best be illustrated by example. The fragment 3,4-dihydroxyphenethylamine (Figure 9) seems to be the moiety responsible for the anti-Parkinson activity of L-Dopa. Consequently, it was of interest to chemists and biologists to ascertain whether other compounds in our files contained this structural unit. Such a search is virtually impossible using rotated lists of WLN's.

For this search, the aromaticity of the ring was not considered so that the query would be more generic. The numbering of the fragment can be random when creating the query matrix. Each query atom is assigned a cyclic qualifier as is the case in the generated matrix. However, the cyclic qualifier for a query atom can be more generic than those

Atom Number	Symbol	Neighbors (Connection)	Cyclic Qualifier
1	Y	2 9 12	2
2	S	1 3 13	2
3	Y	2 4 7	2
4	L	3 5	2
5	L	4 6	2
6	T	5 7 14	2
7	Y	3 6 8	2
8	Y	7 9 15	2
9	Y	1 8 10	2
10	L	9 11	2
11	L	10 12	2
Segment A 12	L	1 11	2
13	W	2	0
14	M	6	0
15	L	8 16	0
Segment B 16	L	15 18	0
17	M	18 21	2
18	Y	17 19 16	2
19	L	18 20	2
20	L	19 21	2
Segment C 21	L	17 20	2

Figure 8.



3,4-dihydroxyphenethylamine

Figure 9.

Cyclic Classification	Qualifier
Acyclic	0
Any type	1
Any ring (including phenyl)	2
Non-phenyl ring	3
Phenyl ring	4

Figure 10.

assigned in a generated matrix (Figure 10). The complete query matrix is shown in Figure 11.

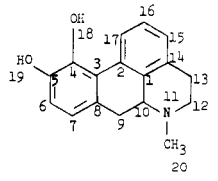
Atom numbers 1 through 9 illustrate the generic range of the query matrix. For example, atom number 2 is a carbon atom which may have any one of six different bonding patterns (see Figure 5). On the other hand, the query may be highly specific such as requiring the oxygen functions (atoms 10, 11) to be nothing other than hydroxyl groups.

When the 3,4-dihydroxyphenethylamine fragment search was performed, one of the compounds found was apomorphine (Figure 12) which will be used to describe the search algorithm. The matrix generated for apomorphine is shown in Figure 13.

The first step of the search is to build "sets" of target atom numbers for each query atom by comparing query atoms (Figure 11) with target atoms (Figure 13). In order for a target atom to be added to a set, it must match the query atom in DPS and cyclic qualifier. For example, the symbols N and M of query atom 1 are compared with the symbols of target atoms 1–10 without matches. A match, including the cyclic qualifier, is found between symbol N, query atom 1 and target

Atom Number	DPS	Neighbors (Connections)	Qualifier
1	N,M	2	1
2	C,D,L,T,X,Y	1 3	1
3		2 4	1
4		3 5 9	2
5		4 6	2
6		5 7 10	2
7		6 8 11	2
8		7 9	2
9		4 8	2
10	Q	6	0
11	Q	7	0

Figure 11.



Apomorphine

Figure 12.

Atom Number	DPS	Neighbors (Connection)	Cyclic Qualifier
1	T	2 10 14	2
2	T	1 3 17	2
3	T	2 4 8	2
4	T	3 5 18	2
5	T	4 6 19	2
6	D	5 7	2
7	D	6 8	2
8	T	3 7 9	2
9	L	8 10	2
10	Y	1 9 11	2
11	N	10 12 20	2
12	L	11 13	2
13	L	12 14	2
14	T	1 13 15	2
15	D	14 16	2
16	D	15 17	2
17	D	2 16	2
18	Q	4	0
19	Q	5	0
20	1	11	0

Figure 13.

atom 11. Target atom 11 is added to the set of query atom 1. No comparison is made between symbol M, query atom 1 and target atom 11 since the target atom had already been included in the set. The comparisons between query atom 1 and the rest of the target matrix are continued with no further matches.

Query atom 2 is then compared with the entire target matrix, and, because of the generic nature of the carbon atom, nearly all the target atom numbers are added to the set for query atom 2.

This process is repeated until every query atom has been compared with every target atom. Each query atom must have at least one target atom match. If at any point a null query atom set results, no further matches are necessary as the target

Query atom	Target atom numbers
1	11
2-9	1-10, 12-17
10	18, 19
11	18, 19

Figure 14.

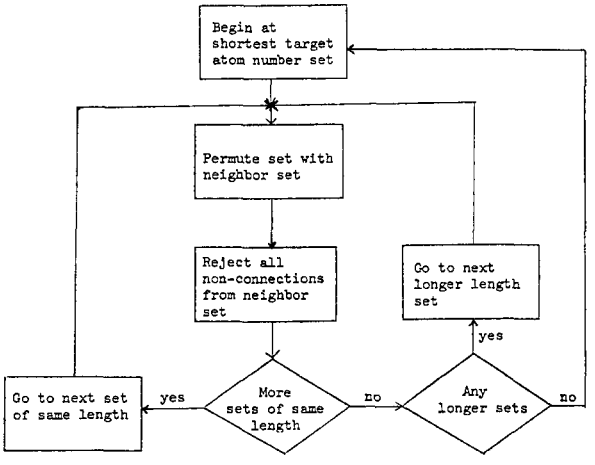


Figure 15.

Query atom	Target atom number
1	11
2	10 12
3	1 9
4	2 8
5	3
6	4
7	5
8	4 6
9	3 7
10	18
11	19

Figure 16.

compound cannot satisfy the query. The complete sets generated before initiating the search are shown in Figure 14.

The search algorithm (Figure 15) may be considered a path-tracing medium which examines many paths simultaneously and obviates the need for backtracking. This is accomplished by a series of reductions of the target atom number sets. For instance, the query matrix has a connection between atoms 1 and 2. Therefore, the target matrix must show a connection between the target atom numbers of sets 1 and 2. Set 1 contains only target atom 11. Set 2 contains 16 target atom numbers, most of which obviously cannot be connected to target atom 11. Those which are not connected are eliminated from the neighbor set. Both sets are permuted to test all combinations. This results in the retention of atom numbers 10, 12, and 20 in set 2. This comparison is performed on each set of target atom numbers.

Searching results in a situation where continued scanning of sets produces no further reduction (Figure 16). Although the presence of the desired isomorph can be seen in the remaining structure (Figure 17), positive proof has not yet been established because of the presence of extraneous atoms 1 and 2.

In such a case, further set reduction is achieved by selecting the last member of the shortest set containing more than one member; a new set containing all the members of the earlier

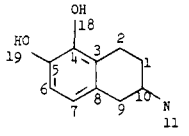


Figure 17.

Query atom	Target atom number	
	Old Set	New Set
1	11	11
2	10	12
3	1 9	1 9
4	2 8	2 8
5	3	3
6	4	4
7	5	5
8	4 6	4 6
9	3 7	3 7
10	18	18
11	19	19

Figure 18.

Table I

	No cyclic qualifier	Cyclic qualifier
Records processed	82 021	82 021
Records searched	63 929	47 373
Search time (min)	113	97
No. of set atoms	5.1×10^6	3.4×10^6
No. of tests	27.7×10^6	17×10^6
No. of loads	9.2×10^4	5.2×10^4

sets and the single member of the selected set is generated. The single selected member from the earlier set is erased (Figure 18).

The reduction technique is applied to the new set, which eliminates the superfluous atoms and retains those satisfying the search requirements. A hit occurs when the number of target atoms agrees with the number of query atoms and no null or duplicate sets exist. If any other conditions exist, the substructure is not part of the target molecule. Apomorphine is a hit because all search criteria are satisfied. This compound has anti-Parkinson activity.

The data presented in Table I are a compilation of several "real-world" searches, which were submitted to our Research Records Office by members of the chemical and biological staffs. First, relevant notation subfiles were created by using bit and string search techniques which generally reduced the file by 80 to 90%. Atom-by-atom searching followed. The notations in each subfile were reviewed to establish the validity of the search algorithm. No failure of the algorithm was recorded. Searches were run with and without a cyclic qualifier to determine the value, if any, of its use.

Running searches against the same subfiles give the figures in Table I. The number of set atoms indicates the total number of target atom numbers which were incorporated into the query sets before initializing the search algorithm. A one-third reduction has been realized from the use of the cyclic qualifier. The number of tests and comparisons necessary to establish whether the compound in question satisfied the query was reduced by almost 40% using the cyclic qualifier.

Finally, the number of loads, too, was greatly reduced by utilizing the cyclic qualifier. "Loads" occur when the search process reaches the point where no further reduction is possible and no solution has yet been obtained. At this point a "new" set is generated, each target atom number being a "load". From these results it is obvious that the use of a cyclic qualifier

has contributed significantly to the efficiency of the search algorithm.

Atom-by-atom searching is done on-line utilizing a Honeywell H-6080 computer and a terminal. On-line searching has enabled us to alter queries, based on initial results, and research the same file without being dependent on turn-around time. On the other hand, should the search file contain more than 5000 WLN's, the search can be done in a batch environment. The average search, which includes bit and string screening, matrix generation, and atom-by-atom searching, consumes approximately 10 min of processing time.

Since, for our purpose, it is not important to analyze polymers, chelates, and inorganics, the system rejects these classes of compounds. Procedures for processing WLN's which contain nonconsecutive locants have been worked out and await implementation into the system.

At present, research continues with the aim of extending the capability, increasing the efficiency, and utilizing the programs to solve related problems. The use of the internal file of over 120000 WLN's has given us an excellent testing ground for our programs, which took approximately 2.5 man-years to develop. The economics of utilizing the programs on the ICRS file of approximately 1.5 million structures are also being examined.

ACKNOWLEDGMENT

We wish to acknowledge Mrs. Olga Z. Buchko for her efforts in the preparation of this paper.

REFERENCES AND NOTES

(1) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(2) P. F. Sorter, C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, "Rapid Structure Searches via Permuted Chemical Line-Notation", *J. Chem. Doc.*, **4**, 56–60 (1964).
(3) A. Sheng, L. Lupi, M. Ronayne, A. Sprules, and S. Zornetzer, "Hoffmann-La Roche's On-Line/Batch Interactive Chemical Information System", *J. Chem. Doc.*, **14**, 179–185 (1974).
(4) E. G. Smith, and P. A. Baker, "The Wiswesser Line-Formula Chemical Notation (WLN)", CIMI, Cherry Hill, N.J., 1975.
(5) E. Hyde, F. W. Mathews, L. H. Thomson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds", *J. Chem. Doc.*, **4**, 200–209 (1967).
(6) W. J. Wiswesser, "Punched-Card Pictures of Atomic Arrangements", *KWIK List News* (Feb 14, 1964).
(7) W. J. Wiswesser, "Eightfold Key to Line-Formula Symbols", *KWIK List News* (March 17, 1964).
(8) E. H. Sussenguth, Jr., "A Graph-Theorectic Algorithm for Matching Chemical Structures", *J. Chem. Doc.*, **5**, 36–43 (1965).
(9) The method of analysis is similar to that developed by C. E. Granito, S. Roberts, and G. W. Gibson, "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems", *J. Chem. Doc.*, **12**, 190–196 (1972).

Investigation of the Index Structure of Drugdoc and Ringdoc†

RIHEI FUJIMOTO

Dainippon Pharmaceutical Co., Ltd., Enoki 33-94, 564 Suita, Osaka, Japan

Received March 15, 1976

Drugdoc, a computerized information retrieval system, is part of the Excerpta Medica Automated Storage and Retrieval Program of Biomedical Information which provides data from about 280000 biomedical research papers per year taken from about 3500 medical journals. The Drugdoc input comprises drug-related information selected from the total data base, augmented by similar information gathered from 200 journals pertaining to pharmaceutical science, a total of 50000 reports per year.

In 1974, a study committee[1] published a comparative study on the input data base, system construction, and characteristics of the Drug Literature Index and Adverse Reactions Titles which can be regarded as the hard copy of Drugdoc.

Ringdoc, a drug information system run on a membership basis, provides 40000 abstracts yearly of articles taken from about 360 drug-related core journals and is published as an abstract journal and microfilm. Also, for retrieval purposes, index cards and punched cards or magnetic tapes are provided.

Smith et al.[2] and Ashmole et al.[3] have reported comparative studies on the retrieval ratio and cost effectiveness of Excerpta Medica, Ringdoc, and other systems. The Drugdoc Study Committee[4] also reported a comparative study of Drugdoc and Ringdoc from the coverage of both systems in 14 drug-related biomedical journals and on the output and distribution of subject areas according to importance to the drug propranolol.

This paper presents the results of a comparative study on the coding and consequential index structure between Drugdoc and Ringdoc on three drugs to ascertain the difference in selection of original articles between the two systems.

RETRIEVAL PROCEDURE

(1) From the Drugdoc S.D.I. output and Ringdoc abstracts of 1974, articles appearing in both systems and related to either of three drugs, the antidepressant Nortriptyline, the β-adrenergic blocking agent Propranolol, and the antibiotic Erythromycin, were selected. From these 60 articles, one article for each drug in the fields of microbiology or pharmacology, therapeutics, and toxicology was further selected, for a total of nine articles. The bibliographies of these are shown in Table I. The Drugdoc S.D.I. output and the Ringdoc abstract for one of the articles are shown in Tables II and III.

(2) In Drugdoc, three types of descriptors are used for coding pertinent material, Preferred Term, Secondary Indexing Term, and Item Index, these being listed in the computer output of Drugdoc. As the Item Index is a numeral code, in this paper it is converted to its related word.

The descriptors in Ringdoc are Index Term, Free Term, Thematic Group Code, and S.D.I. Profile Code, which together make up the Codeless Scanning heading for the abstract. The Codeless Scanning heading for the abstract in Table III is the first seven lines above the journal citation, the Index Terms being underlined, the Thematic Group Code's letters to the left of the heading, and the S.D.I. Profile Code given in Roman numerals.

The weekly output of Ringdoc includes a Free Term Index, comprising the more important free terms for manual retrieval. In this paper Free Term designates free terms in this index

† It should be stated that in Ringdoc there are two coding systems, which are coded independently by different groups of coders and are both available for information retrieval. One is the Ringcode system which is a fragmentation code originally developed for the punched card and the other the Codeless Scanning system explained in this paper. The author has deliberately ignored the former system in this study in order to compare Drugdoc and Ringdoc on an identical basis.

This paper was presented at the 10th National Convention for the Study on Information and Documentation, Japan Information Center for Science and Technology (in Japanese), Oct 15, 1974 (Tokyo, Japan).