

a grant from the Office for Scientific and Technical Information (UK). The grant holder was Sir Ewart Jones, F.R.S., Professor of Organic Chemistry at the University of Oxford. Notwithstanding these acknowledgments the authors take sole responsibility for the contents of the paper.

LITERATURE CITED

- (1) Crowe, J. E., Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "The Searching of Wiswesser Line Notation by Means of a Character-Matching Serial Search," *J. Chem. Doc.* **13**, 85-92 (1973).
- (2) Lancaster, F. W., "MEDLARS: Report on the Evaluation of its Operating Efficiency," *Amer. Doc.* **20**, 119-42 (1969).
- (3) Crowe, J. E., Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "The Development and Evaluation of a Current Awareness Service Based on the *Index Chemicus Registry System*," The Experimental Information Unit, University of Oxford, June 1973. Report to the Office for Scientific and Technical Information (U.K.).
- (4) Leggate, P., "The Evaluation of Operational Current Awareness Services. A Discussion of Practical and Theoretical Problems of Experimental Design," The Experimental Information Unit, University of Oxford, September 1971. OSTI Rept. No. 5111.
- (5) Leggate, P., Smith, B., Stow, J., Williams, M. I., "The *BA Previews* Project: the Development and Evaluation of a Mechanised SDI Service for Biologists," The Experimental Information Unit, University of Oxford, January 1973. OSTI Rept. No. 5139.
- (6) Wilcoxon, F., "Individual Comparisons by Ranking Methods," *Biometrics Bull.* **1** (1), 80-3 (1945); and "Probability Tables for Individual Comparisons by Ranking Methods," *Biometrics* **3** (3), 119-22 (1947).
- (7) Garfield, E., Editorial in *Current Abstracts of Chemistry and Index Chemicus* **40** (1) (6 January 1971).
- (8) Corfield, M. G., Firth, R. J., Fraser, G., Hartley, D., Leggate, P., Norgett, M. M., Riley, J., Rossiter, B. N., "The Liaison Scientist Experiment: A Study of the Provision of Mechanised Current Awareness Services to University Chemists," The Experimental Information Unit, University of Oxford, May 1973. OSTI Rept. No. 5169.
- (9) Rubinstein, R. I., and Qazi, A., "Alternatives to Searching Semantic Surrogates of Chemical Structures," *J. Chem. Doc.* **11**, 110-16 (1971).
- (10) Garfield, E., "Introducing ANSA—ISI's Automatic New Structure Alert—A Compound-Retrieval Service for People more interested in Compounds than Retrieval," *Current Contents* **13** (19), 5-10 (9 May 1973).

Use of the Sequence Rule for Indexing Functional Groups in Organic Compounds

PAUL F. HUDRLIK

School of Chemistry, Rutgers University, The State University of New Jersey, New Brunswick, N. J. 08903

Received September 26, 1973

A new method of indexing functional groups in organic compounds is described, utilizing the Cahn-Ingold-Prelog sequence rule. Functional carbon atoms are first classified by "functionality," a measure of the oxidation state, then ordered by means of a modified sequence rule. Substructure searching and other applications are discussed.

Searches for organic compounds containing a particular subunit or substructure are frequently required of organic chemists. Conventional indexes are not well suited for such searches.¹⁻¹¹ This paper describes a new method of indexing functional groups in organic compounds, based on an index of functional carbon atoms, which should facilitate such searches. The ordering of the index is based on atomic number and oxidation state, thus related functional groups are near one another and the system is easy for the average chemist to use.

BACKGROUND

An organic chemist usually looks for information concerning (1) a specific compound or compounds, (2) a class of related compounds—i.e., compounds containing a given subunit or substructure, or (3) a chemical reaction.⁸

Searching for information about a specific compound is usually not difficult with conventional indexes. Particularly with formula indexes, a given compound will have a

unique, predictable place. (In formula indexes, where a number of compounds have a common molecular formula, the addition of a structural formula would be desirable.⁸ The use of linear notations³ or structure codes^{1,3,12} for this purpose has been suggested.) Searching for information about substructures or about reactions (other than "name" reactions) is much more difficult.

Since many substructures contain functional groups, and since functional groups are intimately involved in most chemical reactions, the development of a logical index of functional groups may ultimately facilitate reaction searching as well as substructure searching.

A system for classifying or indexing functional groups should satisfy three criteria:²

(1) Any functional group should be listed in one unique, predictable place

(2) Related functional groups should be located near one another, as much as possible

(3) The system should be easy to understand and use

Traditional indexes are easy to use but do not meet the first two criteria. Conventional formula indexes and ring indexes are completely unsuitable for functional group

searches. Permuted formula indexes^{3,10,13} or rotated formula indexes⁶ may have some utility for substructures containing unusual elements, but are not generally satisfactory for functional group searching.

Subject and keyword indexes also have a number of disadvantages, since a given functional group can be listed under more than one name, and since related functional groups will in general not be found together. To the extent that rules are formulated to circumvent these disadvantages, use of the system will require a knowledge of these rules, and it may be less convenient to use. Such a compromise between the first two criteria and the third will be required of any system based on an alphabetical (or alphanumeric) ordering.^{4,14} The development of indexes of linear notations is a useful compromise which is useful for some substructure searches.^{7,11,15}

Classifications^{2,12,16} and fragmentation codes^{17,18} can be used for some types of substructure searches. It is certainly possible to devise a fragmentation code which will meet the first two criteria, above. However, use of any classification or fragmentation system requires some familiarity with the basis for the arrangement of the system. If this arrangement is arbitrary, the system is not likely to be widely used.

Based on the above considerations, it was felt that a method for indexing or classifying functional groups which was based on chemical principles would have wide utility. The system described here relies on oxidation state and atomic number, two concepts familiar to practicing organic chemists.

THE SEQUENCE RULE

The sequence rule is part of a method developed by Cahn, Ingold, and Prelog to specify the absolute configuration at asymmetric centers in molecules.¹⁹⁻²² Since the sequence rule describes a procedure for listing all of the ligands or groups—i.e., alkyl groups, halogen atoms, functional groups, etc.—attached to the asymmetric center in a unique order of priority, it can be used to develop an index of such ligands. (The term “ligand” is used in a broad sense to mean any singly attached group, to avoid confusion with the term “functional group,” which will be used to refer to any functional portion of an organic molecule, including those with multiple attachments such as double bonds or carbonyl groups.)

A number of representative ligands are tabulated in increasing order of priority in Table I. Briefly, the order is determined by the atomic number of the atom directly attached to the asymmetric center—thus oxygen takes precedence over nitrogen, and nitrogen takes precedence over carbon. Where the atoms attached to the asymmetric center are the same, the priority is determined by the atoms attached to them, and so on. Detailed discussions of the sequence rule, with further examples, are found in the references,¹⁹⁻²³ and the topic is discussed briefly in a number of textbooks.²⁴⁻²⁸

Indexing Substituted Atoms. Just as the sequence rule indicates a method for ranking all ligands in a unique order of priority, the same principles can be used to list all substituted atoms in a unique order, considering all of the attachments to these atoms. Thus, for substituted carbon atoms, the priority is determined by considering the atomic number of the atoms attached to all four bonds rather than three. By this modification, every atom in every conceivable compound could in principle be arranged in a unique order of priority. (In only a very small number of cases would two atoms have exactly the same priority.)

To develop a useful system for classifying and indexing functional groups in organic compounds, two further modifications are made:

Table I. Representative Ligands in Order of Sequence Rule Priority

—H	—NO
—CH ₃	—NO ₂
—CH ₂ CH ₃	—OH
—CH=CH ₂	—OCH ₃
—Ph	—OCOCH ₃
—CN	—OSO ₂ R
—CHO	—OSO ₂ Ar
—COCH ₃	—F
—CO ₂ H	—SH
—NH ₂	—SO ₃ H
—NHCH ₃	—Cl
—NHCOCH ₃	—Br
—NMe ₂	—I

(1) Attention is focused only on carbon atoms in functional groups

(2) These carbon atoms are first classified by *functionality*, then arranged by the sequence rule as modified above. *Functionality*, a simple and convenient measure of the oxidation state at a carbon atom, is defined by Hendrickson²⁹ as the sum of number of bonds to elements more electronegative than carbon and the number of π bonds to carbon. This modification brings together functional groups of the same oxidation state which would otherwise be more scattered—e.g., nitriles, amides, esters, and acid halides (with *functionality* = 3).


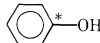





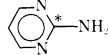







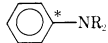

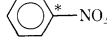

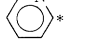
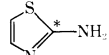

In Table II are listed some representative functional groups, arranged by carbon atom as discussed above. (Some unfunctionalized carbon atoms as in alkanes are included as well.) Next to each formula is a number to facilitate rapid scanning of the list. The first digit is the *functionality* (defined above), and the following pair of digits is the highest atomic number of the elements directly attached to the carbon atom. (This number is a simple fragmentation code. It can be extended into a longer series of digits if necessary, and one can imagine a very long series of digits to specify uniquely the sequence-rule priority of an atom.)

For indexing functional groups in organic compounds, this system has a number of advantages over conventional indexes. Any functional group (including those not yet in existence) will have a unique, predictable place in the index; for the most part, related functional groups (including keto-enol tautomers³⁰) will be found near one another; and the arrangement is based on two concepts which are already well known to the organic chemist—atomic number and oxidation state.

Applications. For use in substructure searching, a polyfunctional compound would be indexed in as many places as there are functional carbon atoms. In principle, every carbon in a molecule could be indexed: however, in many compounds, most of the carbon atoms are not functionalized—i.e., *functionality* = 0. A wide variety of searches would be possible—for example, searches for trifluoromethyl groups, enol ethers, aromatic nitriles, tertiary bromides, aliphatic nitro compounds, or primary sulfonamides would be fairly simple. Some combinations might also be feasible: α -diketones, β -hydroxyaldehydes, *o*-nitrophenols, etc. However, searches for distant combinations (“all aromatic sulfoxides containing a nitro group,”) or for rings (“all cyclobutanes,” “all quinoline derivatives”) would be more difficult.

Textbooks and reference works dealing with synthetic methods and reactions in organic chemistry could include an index of this type (with appropriate modifying phrases such as “synthesis of” and “reaction with—”) in addition to the usual subject and author indexes. Ultimately, the principles discussed in this article should be useful in devising printed indexes of chemical *reactions*, since most organic reactions take place at a limited number of sites

Table II. Index of Functional Carbon Atoms^a

0-01	CH ₄	1-16	RCH ₂ -S-O-R	2-08	CH ₂ = [*] CH-OH	3-08	RCONH ₂
0-03	CH ₃ Li			2-08	CH ₂ = [*] CH-OR	3-08	RCO ₂ H
0-06	RCH ₃			2-08	CH ₂ = [*] CH-OAc	3-08	RCO ₂ Et
0-06	R ₂ CH ₂	1-16	RCH ₂ -S-R	2-08		3-08	RC(OMe) ₃
0-06	R ₄ C			2-08	CH ₂ O	3-09	RCOF
0-12	RCH ₂ MgBr			2-08	CH ₃ [*] CHO	3-09	RCF ₃
1-06	CH ₂ =CH ₂	1-16	RCH ₂ -S-R	2-08	RCHO	3-17	RCOCl
1-06				2-08		3-17	CHCl ₃
1-07	RCH ₂ NH ₂	1-16	RCH ₂ SO ₂ NH ₂	2-08	R ₂ C=O	3-17	CCl ₃
1-07	RCH ₂ NHMe	1-16	RCH ₂ SO ₂ H	2-08		4-07	
1-07	RCH ₂ NHCOCH ₃	1-16	RCH ₂ SO ₂ Cl	2-08			
1-07	RCH ₂ NMe ₂	1-17	RCH ₂ Cl	2-08			
1-07	RCH ₂ NO ₂	1-35	RCH ₂ Br	2-08			
1-08	RCH ₂ OH	1-53	RCH ₂ I	2-16		4-08	NH ₂ -C(=O)-NH ₂
1-08	RCH ₂ OMe	2-06	HC≡CH	2-16			
1-08	RCH ₂ OAc	2-06	CH ₂ = [*] C=CH ₂	2-16		4-08	RO-C(=O)-NH ₂
1-08	RCH ₂ OCOPh	2-07	CH ₂ = [*] CH-NR ₂	2-16		4-08	CO ₂
1-08	RCH ₂ OSO ₂ R	2-07		2-16			
1-08	R ₂ CHOH	2-07		2-17		4-08	RO-C(=O)-OR
1-09	RCH ₂ F	2-07		3-07	R-C≡N	4-16	
1-09	R ₂ CHF	2-07		3-08	R-C≡C-OR	4-16	CS ₂
1-09	R ₃ CF	2-07	R ₂ C=NOH	3-08	CH ₂ = [*] C=O	4-17	CCl ₄
1-16	RCH ₂ SH						
1-16	RCH ₂ SR						

^a Abbreviations: R = n-alkyl, Me = Methyl, Et = ethyl, Ac = Acetyl, Ph = Phenyl. Where several carbon atoms are present the one being indexed is indicated with an asterisk.

which can be conveniently classified by oxidation state and the atomic number of the elements involved.

Indexes based on the sequence rule would also be useful in retrieval of spectral information. An index of substituted carbon atoms as discussed above would be directly applicable to ¹³C NMR spectral information, and similar indexes could be developed for other NMR and IR data.

Just as the above indexes of atoms in molecules are constructed, one can envision indexes based on other centers. For special purposes, one may want an index of substituted double bonds, substituted benzene rings, substituted cyclopropanes, or derivatives of polycyclic ring systems—e.g., steroids.

It is perhaps surprising that systems based on atomic number (or on the periodic table) have not been used in chemical indexing. The system described here, based on the Cahn-Ingold-Prelog sequence rule, should have a number of advantages over conventional indexes for many substructure searches and should have a variety of other applications.

LITERATURE CITED

- Bernier, C. L., "New Kinds of Indexes," *J. Chem. Doc.* 1 (1) 62-7 (1961).
- Buhle, E. L., Hartnell, E. D., Moore, A. M., Wiselogle, L. R., and Wiselogle, F. Y., "A New System for the Classification of Compounds," *J. Chem. Educ.* 23, 375-91 (1946).
- Dyson, G. M., "Studies in Chemical Documentation," *Chem. Ind.* 1952, 676-84.
- Fisher, N. G., "Chemical Indexing: The Literature Chemist's Point of View," *J. Chem. Doc.* 1 (1), 52-6 (1961).
- Frome, J., "Searching Chemical Structures," *Ibid.* 4, 43-5 (1964).
- Garfield, E., "Generic Searching By Use of Rotated Formula Indexes," *Ibid.* 3, 97-103 (1963).
- Granito, C. E., and Rosenberg, M. D., "Chemical Substructure Index (CSI)—A New Research Tool," *Ibid.* 11, 251-6 (1971).
- Loev, B., "Discussion of Some Problems Involved in Using the Chemical Literature," *Ibid.* 1 (2), 27-35 (1961).
- Schmerling, L., "Chemical Indexing: The Research Chemist's Point of View," *Ibid.* 1 (1), 46-51 (1961).
- Skolnik, H., and Hopkins, J. K., "Simplified Stoichiometric Formula Index," *J. Chem. Educ.* 35, 150-2 (1958).
- Skolnik, H., "A Notation Symbol Index for Chemical Compounds," *J. Chem. Doc.* 11, 120-4 (1971).
- Sher, I. H., O'Connor, J., and Garfield, E., "Rotadex—A New Index for Generic Searching of Chemical Compounds," *Ibid.* 4, 49-53 (1964).
- Gelberg, A., Nelson, W., Yee, G. S., and Metcalf, E. A., "A Program for Retrieval of Organic Structure Information Via Punched Cards," *Ibid.* 2, 7-11 (1962).
- Bernier, C. L., and Crane, E. J., "Correlative Indexes. VIII. Subject Indexing vs. Word Indexing," *Ibid.* 2, 117-22 (1962).
- Bonnett, H. T., "Chemical Notations—A Brief Review," *Ibid.* 3, 235-42 (1963).
- Feeman, J. F., "A Novel Organizational Code for Organic Structures Based on Functional Groups," *Ibid.* 6, 184-7 (1966).
- Huber, M. L., "Chemical Structure Codes in Perspective," *Ibid.* 5, 4-8 (1965).
- Maynard, J. T., "A Simplified Chemical Structure Fragmentation System," *Ibid.* 10, 285-9 (1970).
- Cahn, R. S., and Ingold, C. K., "Specification of Configuration About Quadricovalent Asymmetric Atoms," *J. Chem. Soc.*, 1951, 612-22.
- Cahn, R. S., Ingold, C. K., and Prelog, V., "The Specification of Asymmetric Configuration in Organic Chemistry," *Experientia* 12, 81-94 (1956).
- Cahn, R. S., "An Introduction to the Sequence Rule: A System For The Specification Of Absolute Configuration," *J. Chem. Educ.* 41, 116-25, errata 508 (1964).
- Cahn, R. S., Ingold, C., and Prelog, V., "Specification of Molecular Chirality," *Angew. Chem. Int. Ed. Engl.* 5, 385-415, errata 511 (1966).
- "IUPAC Tentative Rules for the Nomenclature of Organic Chemistry. Section E. Fundamental Stereochemistry," *J. Org. Chem.* 35, 2849-67 (1970).

- (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-