

A Notation Symbol Index for Chemical Compounds*

HERMAN SKOLNIK

Hercules Incorporated, Hercules Research Center,[†] Wilmington, Del. 19899

Received September 18, 1970

A new concept is described for the indexing of chemical compounds by the 25 symbols used in a linear notation system introduced recently by the author. The notation symbol index is assigned by listing the notation symbols as they appear in the linear notation, proceeding from the highest numbered to the lowest numbered atom (from left to right), and separating the symbols with a virgule for computer permutation. For example, the notation symbol index for $\text{CH}_3\text{CH}=\text{CHCH}_2\text{OH}$ (linear notation = AB2CQH) is: A/B2/C/QH/ for $\text{CH}_3\text{—/—CH=CH—/—CH}_2\text{—/—OH/}$. In contrast to a formula index, the notation symbol index is unique and discriminatory for each compound and more economical for computer processing than a linear notation permuted index.

Despite the activity and progress in chemical nomenclature, indexing by chemical names is still a challenging problem in chemical documentation systems. Of the various approaches that have been introduced to supplement or complement the chemical name index, the most important are formula and fragmentation indexes and notation and connectivity systems. This paper describes a method, based on a new notation system introduced recently,¹⁰ that yields an index of the radicals associated with each atom in a molecular structure.

THE NOTATION SYSTEM

The uniqueness of the new notation system is the set of notation symbols, shown in Table I, that designates the bonding and number of hydrogens associated with carbon in particular and with other atoms in general. Thus, 13 notation symbols (single letters of the alphabet) define the three important parameters associated with carbon in organic structures—the number of hydrogens attached to a carbon atom, the kinds of bonds on the carbon atom, and the presence of carbon as a fused or bridgehead atom.

Assigning 13 notation symbols to carbon units in a molecule is in harmony with the predominance of carbon and hydrogen atoms in organic chemistry. The high occurrence of the carbonyl group is recognized by assigning to it a separate notation symbol, viz., K. Following carbon and hydrogen, nitrogen occurs most frequently in organic molecules. Consequently three separate notation symbols, M, N, and Z, have been assigned to denote $>\text{NH}$, $>\text{N—}$, and $\equiv\text{N}$ or $=\text{N—}$, respectively.

Chemical structures are represented in the notation system essentially in an atom-by-atom correspondence with the structural formula as usually drawn—i.e., from left to right—with functional groups at the right end position,

or from the atom having the highest position number to the number one atom. This is in accordance with accepted numbering schemes. Figures 1 and 2 illustrate the correspondence between the linear notation and the normal drawing and numbering of two isomeric octenes and three isomeric chlorobenzoic acids, respectively. The one-to-one or atom radical correspondence and conformity with the usual practice of numbering and drawing organic structures have resulted in a notation system unencumbered with inflexible rules.

If a notation system is to be widely useful, it must allow for maximum flexibility with no ambiguity. Flexibility is essential because a chemical structure has many meanings to a chemist. A chemical structure is meaningful as an entity and for the functionalities and moieties, radicals and radical groups, and position relationships within the whole or parts of the molecule. The meaning of structural relationships in a molecule is different for NMR and for reactivity studies. Thus, although in its generalized use, structures are represented in the notation system by proceeding from the highest numbered atom to the lowest, the opposite procedure was used in applying it to a structure—NMR data computer system.¹¹ In applying the notation system for any given use, however, it is essential that consistency be maintained.

Flexibility is inherent also in writing notations. For example, Figures 1 and 2 show two alternatives for writing each linear notation—one with periods and the other with parentheses to denote substituents on a main chain or ring. Each emphasizes a different aspect of the chemical structure.

PERMUTED INDEX

In applying the notation system to correlate proton groups in organic molecules with chemical shifts (NMR data), notations were permuted on the nine notation symbols that represent proton groups, viz., A($\text{CH}_3\text{—}$), B(—CH=), C($\text{—CH}_2\text{—}$), E($=\text{CH}_2$), H, J (bridgehead

* Presented before the Division of Chemical Literature, 160th Meeting, ACS, Chicago, Illinois, September 14, 1970.

[†] Hercules Research Center Contribution Number 1520.

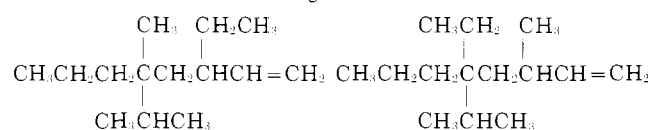
A NOTATION SYMBOL INDEX FOR CHEMICAL COMPOUNDS

>CH—), M(—NH—), U(=CH), and Y(CH—). This is easily and economically done in a computer and yields from one input complete information on all proton groups in each of the chemicals in the file.

Permutation on the 25 notation symbols (Table I) plus several of the character symbols is a relatively simple computer operation. Figure 3 lists 12 C₄H₈O compounds that contain one or more methyl groups (notation symbol is A), with the notation assignments, in alphabetical order relative to the methyl group in a fixed position when permuted on the methyl group. Because two compounds (compounds 1 and 2) have two methyl groups on different atoms, a total of 14 permuted entries results from the 12 compounds. The alphabetization by computer is from left to right on the permuted notation relative to the fixed position.

Permutation on each of the nine different notation sym-

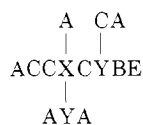
Two-dimensional structural diagrams:



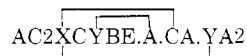
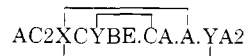
3-ethyl-5-isopropyl-5-methyl-octene

5-ethyl-5-isopropyl-3-methyl-octene

Two-dimensional notational diagrams:



Linear notation (lines drawn to illustrate the attachment):



or

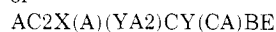


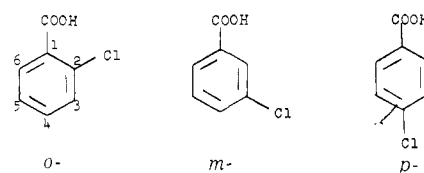
Figure 1. Notations for two isomeric octenes

bols plus the first period of cyclic notations that occur in the 15 C₄H₈O compounds listed in Figure 4 results in a total of 72 notation index entries, or an average of 4.8 per compound.

In permuting notations by computer, a fixed record length for the notations must be delegated for input and storage. The record length, however, must be double that required for the notation to provide for permuting on the fixed position as illustrated in Figure 3. Although relatively few compounds in most files have notations that are up to 40 spaces long, experience dictates that we allow up to 40 spaces for the notation. Consequently, we allocate 80 spaces for the computer record of permuted notations and use position 40 as the fixed point for permutation or wrap-around.

Magnetic tape storage and disk commitments are thus relatively high because of the 80-record length allocation

Two-dimensional structural diagrams:



Linear notations (lines drawn to illustrate the attachment):

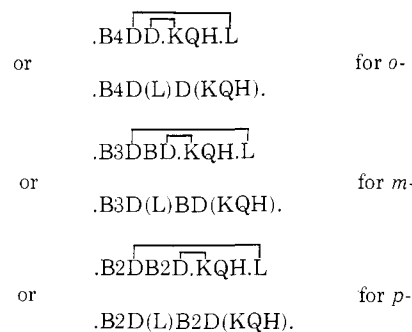



Figure 2. Notations for three isomeric chlorobenzoic acids

Table I. Notation Symbols

Single-Bonded Carbons		Double-Bonded Carbons		Triple-Bonded Carbons		Fused or Bridgehead Carbons			
—CH ₃	A	=CH ₂	E	≡CH	U	>CH—	J		
—CH ₂ —	C	=CH—	B	≡C—	V	>C<	T		
>CH—	Y	=C<	D			=C<	R		
>C<	X	=C=	X						
Carbonyl	Halogen		Oxygen		Nitrogen		Other		
>C=O	K	—F	F	—O—	Q	>NH	M	—H	H
—CH=O	KH	—Br	G	=O	Q	—NH ₂	MH	—S—	S
		—I	I		W	>N—	N	—SH	SH
		—Cl	L	—O—O—	Q2	≡N	Z	=S	S
				—OH	QH	=N—	Z	>SO	SQ
						—N=O	ZQ	>SO ₂	SW
						=NOH	ZQH	P	P
						—NO ₂	ZW		

Character Symbols

- . ~~~~~ cyclic structure
- * fused or bridgehead atom other than C
- # ionic form
- : atoms between bridgeheads or a spiro atom
- condensed ring structure
- ~ polymer repeating unit
- & other atoms + atomic symbol, e.g., &NA for sodium

Structure	Notation	Permuted notation index Fixed Position
1 $\text{CH}_3\text{CH}=\text{CHCH}_2\text{OH}$	AB2CQH	AB2CQH
2 $\text{CH}_3\text{CH}=\text{CHOCH}_3$	AB2QA	AB2QA
3 $\text{CH}_3\text{CH}_2\text{COCH}_3$	ACKA	ACKA
4 $\text{CH}_3\text{CH}_2\text{OCH}=\text{CH}_2$	ACQBE	ACQBE
5 $\text{CH}_3(\text{CH}_2)_2\text{CHO}$	AC2KH	AC2KH
6 $\text{CH}_3\text{OCH}_2\text{CH}=\text{CH}_2$	AQCBE	AQCBE
7 $\text{CH}_3\text{OC}(\text{CH}_3)=\text{CH}_2$	AQD(A)E	AQD(A)E
8 $(\text{CH}_3)_2\text{CHCHO}$	A2YKH	A2YKH
9 $\text{CH}_3\text{CH}-\text{CHCH}_3$.Y(A)Y(A)Q.	.Y(A)Y(A)Q.
(1) $\text{CH}_3\text{CH}_2\text{COCH}_3$	ACKA	ACKA
(2) $\text{CH}_3\text{CH}=\text{CHOCH}_3$	AB2QA	AB2QA
10 $\text{CH}_3\text{OC}(\text{CH}_3)=\text{CH}_2$	AQD(A)E	AQD(A)E
11 $\text{CH}_2-\text{CHCH}_2\text{CH}_3$.CY(CA)Q.	.CY(CA)Q.
12 $\text{CH}_2-\text{CHCH}_3$.C2Y(A)Q.	.C2Y(A)Q.

Figure 3. Permuted notation index of $\text{C}_4\text{H}_8\text{O}$ compounds for the methyl (A) radical

and the resulting high number of permuted entries generated per molecular structure. Whereas the number of permuted entries averaged 4.8 for $\text{C}_4\text{H}_8\text{O}$ compounds, which are relatively small and simple, larger and more complicated molecular structures increase the average permuted entries to approximately 10 per compound.

Formula	Notation
1 $\text{CH}_3\text{CH}_2\text{OCH}=\text{CH}_2$	ACQBE
2 $\text{CH}_3\text{OCH}_2\text{CH}=\text{CH}_2$	AQCBE
3 $\text{CH}_3\text{CH}=\text{CHCH}_2\text{OH}$	AB2CQH
4 $\text{CH}_3(\text{CH}_2)_2\text{CHO}$	AC2KH
5 $\text{CH}_3\text{CH}_2\text{CHCH}_2$.CY(CA)Q.
6 $\text{O}-\text{CHCH}_3$.C2Y(A)Q.
7 $\text{CH}_3\text{CH}=\text{CHOCH}_3$	AB2QA
8 $\text{CH}_3\text{CH}_2\text{COCH}_3$	ACKA
9 $\text{CH}_3\text{OC}(\text{CH}_3)=\text{CH}_2$	AQD(A)E
10 $(\text{CH}_3)_2\text{CHCHO}$	A2YKH
11 $\text{CH}_3\text{CHCHCH}_3$.Y(A)Y(A)Q.
12 $\text{CH}_2=\text{CHCH}_2\text{CH}_2\text{OH}$	EBC2QH
13 $\text{CH}_2\text{CHCH}_2\text{OH}$.C2Y(CQH).
14 CH_2CHOH 	.C3Y(QH).
15 CH_2-CH_2 	.C4Q.

Figure 4. Notation symbol index of $\text{C}_4\text{H}_8\text{O}$ compounds

NOTATION SYMBOL INDEX

Because the stoichiometric formula index is a powerful retrieval mechanism, and has been used extensively to complement chemical subject indexes, the feasibility of basing a similar kind of index on a notation system was investigated. Although the permuted notation system is a partial solution, particularly in its higher discriminatory power relative to the formula index, it is not an economical retrieval mechanism.

The relatively poor discriminatory power of a stoichiometric index is quite apparent by browsing in a *Chemical Abstracts* Formula Index. Thus, in the January-June, 1968 (Volume 68), CA Formula Index, 23 different compounds are listed under $\text{C}_6\text{H}_{14}\text{O}$, 45 under $\text{C}_{10}\text{H}_{14}$, and 57 under $\text{C}_{10}\text{H}_{12}\text{O}$.

There are at least 15 compounds (Figure 4) which can be listed under $\text{C}_4\text{H}_8\text{O}$ (the CA Formula Index, January-June, 1969, listed 13 of these 15).

If we consider the different members of a notation in the same way we do the different atoms in a formula, we obtain a notation symbol index (Column I, Figure 4) that gives considerably more information than does the stoichiometric formula index. The virgule or slanted line between notation symbols is used in our computer programs as a permutation signal. Thus, methyl propenyl ether whose notation is AQB2A, would be in the printout under the following three notation symbol indexes, with only the first as input:

A2/B2/Q/ (input)
B2/Q/A2/
Q/A2/B2/

Notation Symbol Index		
I	II	III
A/B/C/E/Q/	A/C/Q/B/E/	A/C/Q/B/E/
A/B/C/E/Q/	A/Q/C/B/E/	A/Q/C/B/E/
A/B2/C/QH/	A/B2/C/QH/	A/B2/C/QH/
A/C2/KH/	A/C2/KH/	A/C2/KH/
A/C2/Q/Y/	A/C2/Y/Q/	A/C/.C/.Y/.Q/
A/C2/Q/Y/	A/Y/C2/Q/	A/.C2/.Y/.Q
A2/B2/Q/	A2/B2/Q/	A,2/B2/Q/
A2/C/K/	A2/C/K/	A,2/C/K/
A2/D/E/Q/	A2/Q/D/E/	A,2/Q/D/E/
A2/KH/Y/	A2/Y/KH/	A,2/Y/KH/
A2/Q/Y2/	A2/Y2/Q/	A,2/.Y2/.Q/
B/C2/E/QH/	B/E/C2/QH/	B/E/C2/QH/
C3/QH/Y/	QH/C3/Y/	QH/C/.C2/.Y/
C3/QH/Y/	QH/Y/C	QH/.C3/.Y/
C4/Q/	C4/Q/	.C4/.Q/

A NOTATION SYMBOL INDEX FOR CHEMICAL COMPOUNDS

	Formula	Notation	Notation Symbol Index		
			I	II	III
1.		.B3D(A)D(A).	A2/B4/D2/	A2/B4/D2/	A,2/.B4/.D2/
2.		.B3D(A)BD(A).	A2/B4/D2/	A2/B4/D2/	A,2/.B3/.B/.D,2/
3.		.B2D(A)B2D(A).	A2/B4/D2/	A2/B4/D2/	A,2/.B2,2/.D,2/

Figure 5. Notation symbol index of the three xylenes

Regardless of which one of the three we retrieve, it is obvious that the compound is an ether (Q) with a methyl (A) and a propenyl (AB2) radical. This is a lot of information, yet not sufficient to differentiate uniquely each of the 15 C_4H_8O compounds. For example, compounds 1 and 2, 5 and 6, and 13 and 14 in Column I of Figure 4 are not uniquely identified. But each of the other nine compounds is.

The three pairs of duplicate notation symbol indexes arise because Index I (Figure 4) is based on a strictly alphabetical order.

By citing first the atom with the highest number and accumulating similar notation symbols in proceeding to the atom with the lowest number, we obtain the notation symbol index II, Figure 4. Although each of the 15 type II notation symbol indexes is unique, isomers in the aromatic series would not be uniquely identified as illustrated in Figure 5 for the three xylenes. In the case of the xylenes, which is typical of thousands of aromatic isomers, type I and type II have the same single notation symbol index for the three isomers.

Type III, illustrated in Figure 5 for the three xylene isomers, however, does yield a unique notation symbol index for each compound.

The type III index is written as follows:

Write the linear notation, proceeding from the highest to the lowest numbered atom in the molecule, which is generally the order in which the structure is drawn (see Figures 1 and 2).

Proceeding from left to right in the linear notation, list each notation symbol as it occurs with its total occurrence in the molecule. For example, ethyl methyl ketone, whose linear notation is ACKA, starts (at the left) with a methyl group and ends with a methyl group, for a total of two, which is indicated as the first member of the notation index by A,2. On the other hand, a notation symbol contiguous with itself is listed by the notation symbol followed by the total number of contiguous members, such as B4 for the four $-CH=$ and D2 for the two



members of the benzene ring in *o*-xylene (Figure 5).

	Formula	Notation	Notation Symbol Index
1.		.B4D(L)D(A).	L/A/.B3/.D2/
2.		.B3D(L)BD(A).	L/A/.B3/.B/.D,2/
3.		.B2D(L)B2D(A).	L/A/.B2,2/.D,2/
4.		.B3D(L)D(L)D(A).	L,2/A/.B3/.D3/
5.		.B2D(L)BD(L)D(A).	L,2/A/.B2/.D/.B/.D2/
6.		.BD(L)B2D(L)D(A).	L,2/A/.B/.D/.B2/.D2/
7.		.D(L)B3D(L)D(A).	L,2/A/.D/.B3/.D2/
8.		.B2D(L)D(L)BD(A).	L,2/A/.B2/.D2/.B/.D/
9.		.BD(L)BD(L)BD(A).	L,2/A/.B,3/.D,3/

Figure 6. Type III notation symbol index for chlorotoluenes

Notation symbols which represent atoms in a cyclic structure are preceded by a period.

Attachments to a cyclic structure are cited before the ring atoms, citing first the substituent on the highest numbered atom of the ring, as illustrated in Figure 6 for the chlorine and methyl attachments.

The type III notation symbol index yields a unique index for each compound, and, of equal importance, an index from which the formula can be readily drawn. These two advantages are quite obvious in the type III column of Figure 4 for the 15 C_4H_8O compounds.

By including ring atoms as separate entities in the notation symbol index, .B4/.D2/ denotes an *o*-benzene; in a similar manner, .B3/.B/.D,2/ denotes a *m*-benzene, and .B2,2/.D,2/ a *p*-benzene. Figure 6 illustrates the type III index for chlorotoluenes and dichlorotoluenes [note that the substituent on the highest numbered ring atom (C1) is cited first, the methyl group second, and then the members of the benzene ring from position 6 to position 1]. Each notation symbol index is unique for each of the nine compounds in Figure 6, and each discloses unambiguously the molecular structure.

DISCUSSION

The empirical formula arrangements Richter⁷ used in his tables of carbon compounds in 1884 is the first known use of this type of index. A formula index designed by Jacobson and Stelzner⁷ and introduced by Berichte in 1898 was based on the Richter order: C, H, O, N, Cl, Br, I, S, P, and the remaining elements following in alphabetical order. Hill's formula index,⁶ designed for the United States Patent Office in 1900, cited C first, H second, and the other elements thereafter in alphabetical order; CA uses a slightly modified Hill system for its formula index.

Dyson,¹ questioning the value of giving precedence to C and H, developed the "Molform Index" which used the order: P, I, F, Cl, Br, S, N, O, C, H. Fletcher and Dubbs² based their formula index on the periodic table sequence except for C and H, which were cited last. Skolnik and Hopkins⁹ introduced a formula index in which C and H are cited after the other elements which are arranged in alphabetical order.

To provide a generic search tool in *Index Chemicus* for any element contained in a chemical, Garfield³ designed a computer-produced "RotaForm Index" as a by-product. The "RotaForm Index" is basically a permuted formula index for each of the elements.

Granito, et al.^{4,5,12} described a permuted Wiswesser notation index on 47 characters (total of 94) with an average of 6.4 entries per structure by excluding locants, spaces, and other symbols as permuting points.

Permutation of the 15 C_4H_8O notations, without excluding any notation symbol, gives a total of 72 entries for an average of 4.8 per compound. The three isomeric

xylenes have a total of 15 entries for an average of five per compound and the chlorotoluenes have a total of 64 entries for an average of seven per compound. In contrast to permuted Wiswesser notations, the permuted notations illustrated in this paper disclose the presence and relative position of every atom in the molecule.

As already pointed out in this paper, permuting and storage of permuted notations in a computer are relatively costly in terms of CPU time and storage as the record length for permutation needs to be at least twice the length of the longest notation. Computer time and storage are considerably more economical with types I, II, and III notation symbol indexes.

For the 15 C_4H_8O compounds, types I and II notation symbol indexes yield a total of 53 entries or 3.5 per compound and type III a total of 55 entries or 3.7 per compound (Figure 4). For the three xylenes (Figure 5), types I and II average three and type III 3.3 per isomer. For the nine chlorotoluenes (Figure 6, type III), the average is five per compound. The notation symbol index, furthermore, is not restricted by record length, nor are the permutation process and storage unduly costly in computer processing.

REFERENCES

- (1) Dyson, G. M., "Chemical Documentation," *Chem. Ind. (London)*, **1952**, 676-84.
- (2) Fletcher, J. H., and D. S. Dubbs, "Quick Access to Research Records," *Chem. Eng. News*, **34**, 5888-91 (1956).
- (3) Garfield, E., "Generic Searching by Use of Rotated Formula Indexes," *J. Chem. Doc.*, **3**, 97-103 (1963).
- (4) Granito, C. E., A. Gelberg, J. E. Schultz, G. W. Gibson, and E. A. Metcalf, "Rapid Structure Searches via Permuted Chemical Line-Notations. II. A Key-Punch Procedure for the Generation of an Index for a Small File," *J. Chem. Doc.*, **5**, 52-5 (1965).
- (5) Granito, C. E., J. E. Schultz, G. W. Gibson, A. Gelberg, R. J. Williams, and E. A. Metcalf, "Rapid Structure Searches via Permuted Chemical Line-Notations. III. A Computer-Produced Index," *J. Chem. Doc.*, **5**, 229-33 (1965).
- (6) Hill, E. A., "On a System of Indexing Chemical Literature," *J. Amer. Chem. Soc.*, **22**, 478-94 (1900).
- (7) Jacobson, P., and R. Stelzner, "Zur Frage der Benennung und Registrierung der Organischen Verbindungen," *Ber.*, **31**, 3368-88 (1898).
- (8) Richter, M. M., "Tabellen der Kohlenstoff-Verbindungen," R. Oppenheim, Germany, 1884.
- (9) Skolnik, H., and J. K. Hopkins, "Simplified Stoichiometric Formula Index," *J. Chem. Ed.*, **35**, 150-2 (1958).
- (10) Skolnik, H., "A New Linear Notation System Based on Combinations of Carbon and Hydrogen," *J. Heterocyclic Chem.*, **6**, 689-95 (1969).
- (11) Skolnik, H., "A Correlative Notation System for NMR Data," *J. Chem. Doc.*, **10**, 216-20 (1970).
- (12) Sorter, P. F., C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, "Rapid Structure Searches via Permuted Chemical Line-Notations," *J. Chem. Doc.*, **4**, 56-60 (1964).