A New Linear Notation System Based on Combinations of Carbon and Hydrogen

Herman Skolnik

Hercules Research Center, Hercules Incorporated

Dedicated to Professor Allan R. Day

Contribution Number 1492

A new linear notation system is described which was designed to have a one-to-one correspondence with the chemical structures represented by the notations. Each notation is unique and unambiguous, yet simple and compatible with computer input and output characters. The symbols used in the notation system denote carbon in terms of bonds and attached hydrogen(s). The notation system is illustrated with acyclic, carbocyclic, and heterocyclic examples.

Although chemistry is outstanding among the sciences in its literature resource, there are many challenges yet to be solved in our communication and information retrieval mechanisms. For example, in conducting the initial literature searching for five- and six-membered heterocyclics containing oxygen and sulfur and two or more sulfurs (1), the authors were grateful for "The Ring Index" (2) and for the "Index of Ring Systems" in Chemical Abstracts. Yet beyond the one-ring system, almost every page of "The Ring Index" and every entry under the subjects in the Chemical Abstracts Index referred to by the "Index of Ring Systems" had to be scanned. This was a time-consuming, monotonous task.

As the authors delved into the original literature, they encountered many nomenclature problems. They also were confronted with nomenclature problems during the indexing of their book. These problems would have been considerably less burdensome if the heterocyclics they were interested in had been stored in a computer system by some reliable retrieval mechanism.

Ideally a computerized retrieval mechanism for chemical structures should be compatible with accepted practices in the way structural formulas are drawn and in the numbering of chain and ring moieties. It should also be compatible with keypunch input to computers and with the characters of computer print chains. Although many notation systems described in the literature are more or less computer-compatible, their topological delineations are not in harmony with the way structural formulas are drawn or with the accepted numbering, particularly of ring systems. This criterion of compatibility is an important one as there is a high degree of communication in the structural formulas of ring systems. One needs only to examine the cover of *This Journal* to appreciate how much information is transmitted by the structural formula.

In drawing structural formulas, the path is generally from the left to right, but functional groups are assigned the right-end position or as near to the right as possible in acyclics or to the top or right in rings. The numbering, however, generally proceeds from the right, starting with the carbon to which the functional group is attached or is a part of; numbering of the atoms in attachments, however, generally proceeds from the atom attached to the primary moiety. For examples, consider the following compounds:

$$C_6H_5COOCH_2CH=CH_2$$
 allyl benzoate

Fortunately, "The Ring Index" is the bible for ring structures and their numberings. It is oriented to unsaturated ring structures, denoting only saturated atoms with hydrogen.

Basic Notation Principles.

Because structural formulas, as written by chemists, denote combinations of carbon and hydrogen, such as -CH₃, -CH₂-, >CH-, and >C<, it appeared to be logical to base the new notation system on these structural units. Furthermore, the ever increasing importance of nmr in structural and mechanistic studies gives added importance to carbon/hydrogen units.

Table I lists the symbols used in the new notation system for the various combinations of carbon and hydrogen that occur in structural formulas. Carbonyl is included in this group because of its wide occurrence, particularly in ring systems. Table II lists the notations for elements other than carbon.

TABLE 1

Notation Symbols for Combinations of Carbon and Hydrogen and for Carbonyl

Single-bo	onded Carbons	Double-l	oonded Carbons
-CH ₃	A	=CH ₂	E
-CH ₂ -	C	=CH-	В
-ĊH-	Y	>C=	D
-¢-	X	> C=	R (fused >C=)
>CH-	J (fused or bridgehead >CH-)		
><<	T (fused or bridgehead >C<)		
Triple-bo	nded Carbons	Carbony	1
≡СН-	U	>C=O	K
≡C-	V		

TABLE II

Notation Symbols for Atoms Other Than Carbon

Halogen		Nitro	gen	Oxygen	
-F	F	>NH	M	-0-	Q
-Br	G	-NH ₂	MH	-OH	QH
-I	I	>N-	N	O_2 (as in NO_2 or	W SO ₂)
-Cl	L	≡N or -N=	Z		
Other Alphabetical Symbols			Characte	r Symbols	
-H	Н		Fused or bridgehead atom (other than C) is denoted by * following symbol.		
-S- or =S	S	lonic fo symbol.	rm is deno	oted by # foll	owing
P	P	Substitue noted by		n bridgeheads	is de-
		Metal = for sodiu		c Symbol, e.g.,	&NA

In representing a chemical structure, the notations are written as they would be drawn. The numbering system is that of *Chemical Abstracts* or "The Ring Index". Other principles will be noted in discussion of the examples that follow.

Acyclic Hydrocarbons.

The following examples illustrate how the notation system is applied to acyclic saturated hydrocarbons:

Compound	Visual Notation	Written Notation
1. CH ₄		A.H
2. CH ₃ CH ₃	AA	A2
3. CH ₃ CH ₂ CH ₃	ACA	A2C
4. CH ₃ CH ₂ CH ₂ CH ₃	AC_2A	AC2A
5. CHCH ₃ CHCH ₃	A Y A	AYA.A (preferred) A3Y
6. CH ₃ (CH ₂) ₃ CH ₃	AC_3A	AC3A
7. CHCH ₂ CH ₂ CH ₃	A YCA	AYCA.A (preferred) A2YCA
8. (CH ₃) ₃ CCH ₃	A ₃ XA	AXA.A2 (preferred) A4X

In the case of isobutane, isopentane, and neopentane, the preferred notations are: AYA.A, AYCA.A, and AXA.A2, respectively. These conform to IUPAC rules and relate most directly to functional derivatives. The following cases illustrate notations for unsaturated acyclics:

To illustrate the correspondence between the new notation system and the normal writing and numbering of a structure, let us consider 4-ethyl-5-methyl-1-hexane:

It is to be noted that for 5-ethyl-4-methyl-1-hexene, nomenclature rules require the citing of substituents in alphabetical order. This rule is ignored in the notation system, as substituents are more appropriately positioned

in number order; otherwise the position of the substituents would need to be cited. Thus, for the 5-ethyl-4-methyl isomer, the notation is:

Retaining the accepted numbering systems and citing substituents in position number order are basic principles which are compatible with the habits of chemists in writing structural formulas. A consequence of these two principles is the ready delineation of the basic chain or ring (as will be shown later) by a period, which separates the basic chain from its substituents, as in the following:

Only positions 4 and 6 in the basic chain have symbols which denote attachments. Thus, Y requires one attachment which is satisfied by CA; X requires two attachments which are satisfied by A and YA2. This scheme eliminates the need to cite the position of a substituent, which is required by most of the existing notation systems.

A complex branched side chain is denoted as follows:

Notation: AC3YC3A.YAY/A/CA

The slanted line or virgule is used instead of parentheses because it is more common on computer print chains. It is used to show the relationship of a substituent to a branched chain.

Triple bond unsaturation is denoted by U for \equiv CH and V for \equiv C- as illustrated in the following:

CH≡CH	UU	or U2
$CH_3C = CH$		AVU
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		AV2CBE

Acyclic Compounds with Functional Groups.

In harmony with the principle of accepted nomenclature for naming and numbering basic chains, there is no need to establish priorities for functional groups. The following examples illustrate how the new notation system cites functional derivatives:

Compound	Notation
CH ₃ OH	A.QH
CH ₃ CH ₂ OH	AC.QH
СН ₃ СНОНСН ₃	AYA.QH
СН₃СНО	AK.H
CH ₃ COOH	AK.QH
CH ₃ COCH ₃	AKA
CH ₂ =C=O	EK
CH ₂ =CHCH ₂ Cl	EBC.L (L = Cl)
CH ₂ =C(CH ₃)CH ₂ Cl	EDC.L.A
CH ₂ =CHCOOH	EBK.QH
3 CH ₂ =C(CH ₃)COOH	EBK.QH.A
$CH_2 = CHCH_2COOH$	EBCK.QH
$CH_2 = C(CH_3)COOCH_2CH_3$	EDK.QCA.A
CH ₃ CH ₂ CH(NH ₂)COOH	ACYK.QH.MH (M = NH)
N≡CCH ₂ CHCOOH	$VCYK.QH.G.Z.(V = \equiv C.,$
Br	$\begin{array}{cccc} VCYK.OH.G.Z. & (V = \equiv C, \\ & & G = Br, \\ & Z = N \equiv) \end{array}$
$^{3}_{\text{CH}_{3}\text{CH}_{2}\text{CH}_{2}\text{-N-CH}_{2}\text{CH}_{3}}^{2}$	AC2.N/MH/CA
NH ₂	or N.C2A.CA.MH

Nonaromatic Carbocyclic Compounds.

Carbocyclic compounds are delineated under the same basic principles as so far imposed. The order of citation is from the highest to lowest numbered member in the main ring. Fused carbon atoms are denoted by J for >CH-, T for >C<, and R for >C=. The numbering system is that of "The Ring Index". A few illustrative examples are:

Compound	Notation	
Cyclopropane	C3.	
Cyclobutane	C4.	
Cyclopentane	C5.	
Cyclopentene	C3D2.	
Cyclopentadiene 4 1 2	CD4.	
1,2-Dimethylcyclohexane	C4Y2.A2	
1,3-Dimethylcyclohexane	C3YCY.A2	
1,4-Dimethylcyclohexane	C2YC2Y.A2	
Norpinane or Bicyclo[3.1.1] heptane	CJC3J:C.	
6 3 4	$\label{eq:charge} \begin{picture}(J = bridgehead > CH- \ and : = bridgehead \\ alkylene) \end{picture}$	
$\alpha\text{-Pinene}$ or 2,6,6-trimethylbicyclo[3.1.1]hept-2-ene		
H ₃ C 6 7 4	XJCBDJ:C.A3	
Norbornane or Bicyclo[2.2.1]heptane		

654321 7 C2JC2J : C.

Camphene or 2,2-Dimethyl-3-methylenenorbornane

6543 21 7 C2JDXJ:C.A2.E

Camphoi

C2JCKT:X.A.A2 or C2JCKT:X.A3

Aromatic Carbocyclic Compounds.

Aromatic compounds are delineated under the same basic principles as already described. The important symbols used for aromatic compounds are B for =CII-, D for =C<, and R for fused =C< common to two or three aromatic rings. Notation examples are as follows:

Compound	Notation
Benzene	В6.
Benzoic Acid	B5D.KQH
Phthalic Acid.	B4D2.KQH/2
5 COOH	

Toluene B5D.A

p-Toluic Acid

COOH

6

5

4

CH3

Anthracene

7

8

9

1

2

8765 10 9 4-1

B4R2πBR2βπB4.

Phenanthrene

2
3
8-5
84R2=B2R2=B4.

Phenalene

9-7
B3R3nB3R nB2C

8-Chloro-5-(dimethylamino)-1-oxophenalene-2-carboxylic acid
COOH

BDBR3=BDBR=BDK.KQH.NA2.L

BDBR3=BDBR=BDK.KQH.NA2.L

CI 8 7 6 N(CH₃)₂

also drawn

1213

Heterocyclics.

In the following examples, which illustrate the new notation system for heterocyclics, an asterisk following a

mbol denotes a fused ring atom other than carbon (R):			
Ring Index No.	Compound	Notation	
124	Pyrazole 5 NH N ² 4	взим.	
125	3H-Pyrazole 5 N 2 1 CH 2	B2CZ2.	
127	Imidazole	B2ZBM.	
128	2 <i>H</i> -Imidazole N 2 CH ₂ N 3	B2ZCZ.	
881	2 <i>H</i> -Imidazo[4,5- <i>d</i>] oxazole $ \begin{array}{cccccccccccccccccccccccccccccccccc$	MBZRJ¤ŽČQ.	
883	111,3H-Imidazo[1,5-c] oxazole $ \begin{array}{c c} 6 & 7 & 0 \\ 5 & N & 3 \end{array} $	7.65.4 BZBN*R#CQC.	
918	Imidazo [4,5-d] imidazole 5	ZBZR2¤ZBZ.	
919	1 <i>H</i> -Imidazo[1,5-a] imidazole 6N	7654 BZBN*R¤B2M.	
923	1,4-Diazabicyclo [2.2.1] heptane	654 321 7 C2N*C2N*:C.	

6543217 CMJCMJ:M. 925 Trimidine

927 1,3,6,8-Tetraazaspiro[4.4] nonane 9 8 7 6 5 4 3 2 1 CMCM¤T¤CMCM. (¤T¤ = spiro C) 948 Cyclopentimidazole B3R2¤ZBZ.

^{7.4} B4R2¤ZBM,

I-Methylbenzimidazole B4R2¤ZBN.A

 ${\bf 2-Methylbenzimidazole}$ B4R2¤ZDM.A

2-Ethylbenzimidazole B4R2¤ZDM.CA

1-Acetyl-2-methylbenzimidazole B4R2¤ZDN.KA.A сосн3

B3R2¤ZDM.CBE 2-Allylbenzimidazole .CH2CH=CH2

B4R2¤ZDM.BE 2-Vinylbenzimidazole .сн=сн₂

2-Aminobenzimidazole

B4R2¤ZDM.MH

2-Benzylbenzimidazole

R4R2¤ZDM_C/DR5/

4-Chloro-2-chloromethylbenzimidazole

B3DR2¤ZDM.CL.L

5,6-Dimethylbenzimidazole

B2D2R2¤ZBM.A2

2282 8II-Furazano [4,5-e | benzimidazole

MBZR2¤B2R2¤ZQZ.

2730 1H-[1,4]-Oxazino[4,3-a] benzimidazole

B4R2¤N*RZ¤B2QC.

2754 IH-Imidazo[4,5-b] quinoline

B4R2¤ZR2B¤ZBM.

2785 1H-Naphth[2,3-d]imidazole

B4R2¤BR2B¤ZBM

2788 1H-Naphth[1,2-d]imidazole

B4R2¤B2R2¤ZBM.

Discussion.

Chemical notation systems have been with us since the 1940's (3, 6), yet none really has caught the fancy of bench chemists. Nevertheless, the Dyson (4) and the Wiswesser (5) systems have experienced some acceptance by literature chemists for chemical indexes. This acceptance, even though limited, indicates that linear notation systems can contribute to the solution of problems in nomenclature, indexing, and communication vis-a-vis the computer.

The basic weakness of notation systems has been the complexity of the many rules invoked to ensure that a unique and unambiguous notation is assigned to a given structure. These rules have forced upon the potential users the need to learn a different way of visualizing chemical structures and of numbering chains and rings. In designing the notation system described in this paper, the primary objectives were to retain accepted numbering schemes, to have the notations conform to accepted chemical structures, and to invoke relatively few rules. These three objectives have been achieved.

Nine notation symbols define the two important parameters associated with carbon in organic chemical structures, viz., the carbon bonding and the number of hydrogens attached to the carbon. These nine notation symbols are: A (-CH₃), B (-CH=), C (-CH₂-), D (>C=), E (=CH₂), U (=CH), V (=C-), X (>C<), and Y (>CH-). An additional three notation symbols are used to define carbon in condensed rings or as bridgeheads: J (>CH-), R (>C=), and T (>C<); the sumbol μ indicates a condensed ring situation, and μ T μ a spiro carbon; the punctuation symbol: indicates a substituent between bridgehead atoms. The notation symbol K is used to describe the carbonyl function.

Denoting carbon in terms of bonds and hydrogens is a logical approach as these units comprise the major part of organic chemicals. Because other systems tend to treat carbon as acyclic and carbocyclic, they must use symbols to indicate carbon bonding other than single bonding (although some systems, such as atom connection tables, also indicate single bonding). Because the bonding in the new system is included in the symbol, molecular formulas and molecular weights can be calculated easily with a computer and given in the printout along with other information.

Because the notation system described in this paper conforms to practice, visual inspection of a notation reveals more structural information than do other notation systems. With a little experience, one soon is able to project the linear notation into a two dimensional structure, pretty much in the following manner for 2-ethylbenzimidazole:

B4R2¤ZDM.CA

This one-to-one correspondence between a notation system and the structure it represents is a major advantage. At the same time, the notation is unique and unambiguous as a communication. The symbols used are simple, can be inputed by keypunch or typewriter, and are on most computer print chains. It is particularly suitable for computer processing, and requires relatively simple programming for manipulating in the computer from various viewpoints. Its use in computer operations, however, is a subject for a subsequent paper.

REFERENCES

- (1) D. S. Breslow and H. Skolnik, "Multi-Sulfur and Sulfur and Oxygen Five- and Six-Membered Heterocyclics", Volume 21 (in two parts) of "The Chemistry of Heterocyclic Compounds", John Wiley & Sons, Inc., 1966.
- (2) A. M. Patterson, L. T. Capell, and D. F. Walker, "The Ring Index", 2nd Edition, American Chemical Society, 1960.
- (3) "Survey of Chemical Notation Systems", Publication 1150, National Academy of Sciences/National Research Council, 1964.
- (4) "Rules for I.U.P.A.C. Notation for Organic Compounds",
- John Wiley & Sons, Inc., 1961.
 (5) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill Book Co., 1968.
 - (6) Other references to notation systems are:
- (a) M. M. Berry and J. W. Perry, Chem. Eng. News, 30, 407 (1952). (b) R. Fugmann, Nachr. Dokumentation, 12, 69 (1961). (c) W. Gruber, Angew. Chem., 61, 429 (1949). (d) H. W. Hayward, Patent Office Research and Development Reports, No. 21, 1961. (e) J. A. Silk, J. Chem. Doc., 1, 58 (1961). (f) H. Skolnik and A. Clow, ibid., 4, 221 (1964).

Received July 30, 1969

Wilmington, Delaware 19899