Yet Another Linear Notation Scheme for Organic Compounds. 1

N. GHOSHAL

Indian Institute of Chemical Biology, Calcutta-700032, India

Received April 4, 1990

An extended version of the Wilcox and Levinson linear notation scheme for computer representation of chemical structures has been proposed. Use of the proposed EWL (Extended Wilcox and Levinson) notation has been illustrated with examples. Features of this linear representation scheme have been compared with that of the alternatives. It has been shown that all the advantages of the original (Wilcox and Levinson) method have been retained in the extended scheme. Further, the new scheme allows compact and "natural" encoding of substituted compounds, compounds with charged atoms, and aromatic compounds involving a lone-pair of electrons. Syntactic consistency and semantic consistency of the proposed scheme have been verified with the help of computer programs.

INTRODUCTION

Advantages of linear notation[1-3] as an alternative approach for computer representation of organic compounds are fairly well-known. The earliest form of linear notation is the WLN.[5] Though it still finds its adherants, WLN methodology involves a large set of complex rules[2,7] and is said to be a close kin of cryptography. Newer linear notation schemes,[6-9] however, put more emphasis on simple encoding rules and on clarity of the encoded expressions. Consequently, these newer schemes are easier to learn. Two such schemes are the SMILES system[7] of Pomona College and the scheme proposed by Wilcox and Levinson.[8] The latter is hereafter called WLS for convenience.

Expressions in SMILES, on the average, are more compact due to single and aromatic bond suppression. This, however, adversely affects readability. The encoding methodology is fairly simple for acyclic compounds, but for polycyclic compounds SMILES methodology is rather involved.[7] Readability of SMILES expressions is further affected by the nesting of parentheses which are used to indicate branching.

Though not as well-known as SMILES, WLS is simpler to use and encode. WLS requires that the structure to be represented be first dissected into linear and/or monocyclic segments and then each such segment be separately encoded. In many cases this corresponds closely to a chemist's natural perception of a structure. (CONOL-II,[6] another linear notation, also possesses this naturalness in a limited manner.) Notations for segments are separated by commas. This segmentation leads to natural breaks which help the user during computer entry and also during columnar display (see Table I for examples). In WLS, compactness may be achieved by the use of abbreviated forms of some commonly used structures. A brief note on WLS encoding method is given in the Appendix.

However, WLS has the following limitations:

- Necessity to use a large number of segments for certain substituted compounds, e.g., hexachlorocyclohexane is represented as (C1–C2–C3–C4–C5–C6–), C1–Cl, C2–Cl, C3–Cl, C4–Cl, C5–Cl, C6–Cl. This is inconvenient and not natural in the sense described above.
- Absence of provision for certain kinds of bonds, viz., coordination bonds, and distinction between delocalized and aromatic bonds etc. These types of bonds are required to represent many medicinally interesting compounds. In particular, prediction of physicochemical properties[2] like lipophilicity/hydrophobicity, partition coefficient,[10] and molar refractivity,[10] by additivity methods requires that bond types be differentiated into more varieties than those allowed in WLS.
- The intramolecular formal charge distribution (which is required for a Cambridge-type connection table[4,11]) can not be predicted.
- In some aromatic rings involving lone-pair contribution, e.g., pyrrole, hydrogen computation is difficult.

In this paper an extension of the WLS method is proposed whereby the shortcomings of the original scheme are eliminated without sacrificing its simplicity. Attempts have also been made to maintain downward compatibility with WLS, i.e., notations which are valid in WLS would also be valid in the extended scheme, except for compounds with delocalized bonds.

THE PROPOSED REPRESENTATION SCHEME

The extended system, hereafter called EWL, offers the following advantages:

- Substituted compounds can be expressed more compactly, through the use of dangler bond symbols.
- More types of bonds can be represented.
- Charged atoms can be specified.
- Aromatic compounds can be unambiguously represented, as the rules for encoding them have been more precisely defined.

A detailed description follows.

Dangler Bonds. When a side chain (in the hydrogen-suppressed graph) contains exactly one heteroatom and/or a methyl group, dangler bond symbols may be used to economize. To start with we will consider a single dangler bond denoted by (′), e.g., hexachlorocyclohexane would be represented as (C′Cl–C′Cl–C′Cl–C′Cl–C′Cl–C′Cl–) and 2-methyl-3-chloropentane would be represented as C–C′C–C′Cl–C–C.

Side-chain double bonds are denoted by (″), e.g., ethyl methyl ketone is represented as C–C–C″O–C. More than one dangler bonds may be attached to the same atom symbol, e.g., 2,2-dichloropropane would be represented as C–C′Cl′Cl–C and 2-chloro-2-methylbutane as C–C′Cl′C–C–C.

Distinction between Aromatic Bond and Delocalized Double Bond. In EWL the aromatic bonds and the delocalized double bonds are represented by two different symbols which are (+) and (~) respectively, e.g., chlorobenzene would be represented in EWL by (C+C′Cl+C+C+C+C+) and butadiene in EWL by C~C~C~C. The distinction between the bonds as shown above also conforms to the Cambridge connection table convention. In SMILES there is no special symbol for a delocalized bond.

However, there is more to aromaticity than the simplistic concept of aromatic bonds. A discussion of the finer points is deferred to a subsequent section. Ghosh and Crippen[10] make

0095-2338/90/1630-0308$02.50/0 © 1990 American Chemical Society

Table I. Representation of Some Organic Compounds in Linear Notation

compound	WLS	EWL	SMILES
mucic acid (I)	O=C1-C2-C3-C4-C5-C6=O, C1-O, C2-O, C3-O, C4-O, C5-O, C6-O	O=C'O-C'O-C'O-C'O-C'O-C'O=O	OC(=O)C(O)C(O)C(O)C(O)C-(=O)O
betaine hydrochloride (II)	not possible	C-N'Q'C'C-C-C''O-O, Cl'E	C[N+].[Cl-](C)(C)CC(=O)O
hexachlorobenzene (III)	(C1+C2+C3+C4+C5+C6+), C1-Cl, C2-Cl, C3-Cl, C4-Cl, C5-Cl, C6-Cl	(C'Cl+C'Cl+C'Cl+C'Cl+C'Cl+C'Cl+)	c1(Cl)c(Cl)c(Cl)c(Cl)c(Cl)c1(Cl)
glucosamine (IV)	(O-C1-C2-C3-C4-C5-), C1-C-O, C2-O, C3-O, C4-N, C5-O	(O-C1-C'O-C'O-C'N-C'O-), C1-C-O	O1C(CO)C(O)C(O)C(N)C1(O)
biphenyl (V)	(C1+C+C+C+C+C+), (C2+C+C+C+C+C+), C1-C2	(C1+C+C+C+C+C+), (C2+C+C+C+C+C+), C1-C2	c1ccccc1c2ccccc2
pyrrole (VI)	?	(C+C+n+C+C+)	Hn1cccc1 or [H]n1cccc1 or N1C=CC=C1
folic acid (VII)	(C1+N+C2+N+C3+C4+), C4+N+C5+C+N+C3, (C6+C+C+C7+C+C+), C7-C8-N-C9-C-C-C10-O, C5-C-N-C6, C8=O, C9-C11-O, C11=O, C10=O, C1-O, C2-N	(C'O+N+C'N+N+C1+C2+), C2+N+C3+C+N+C1, (C4+C+C+C5+C+C+), C5-C''O-N-C6-C-C-C''O-O, C3-C-N-C4, C6-C''O-O	n1c2c(O)nc(N)nc2ncc1CNc3ccc-(cc3)C(=O)NC(C(=O)O)CCC-(=O)O

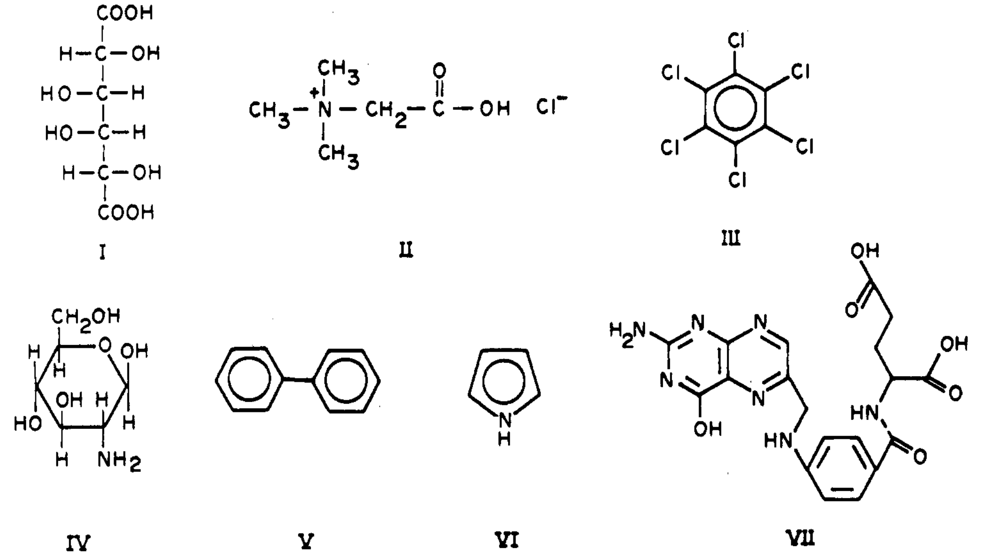


Figure 1. Structures of compounds given in Table I.

the distinction between pyrrole-like and pyridine-like C–N aromatic bonds. It will be shown how the distinction can be retained without introducing new bond symbols.

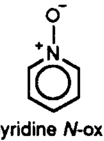
Formal Charge. Ionic bonds between two atoms are often conceived as opposite formal charges between two components having no covalent link. Coordination bonds on the other hand are considered as a combination of formal charges and covalent bonds. Isolated ions require that their formal charges and positions be appropriately shown. The Cambridge connection table has a provision for representing formal charges, SMILES has similar provisions, but WLS does not.

In EWL, formal charges on atoms are represented by attaching the symbol Q for a positive charge and E for a negative charge. The attachment is done by using the dangler bond symbol ('). Thus the syntax of the linear notation remains intact and no special connective is required, e.g., betaine hydrochloride (II of Figure 1) may be represented as C-N'Q'C'C-C-C''O-O, Cl'E.

For semipolar (coordination) bonds, the structure may be represented by a covalent bond with opposite charges on the participating atoms, e.g.

$^{-}C\equiv O^{+}$

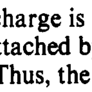
carbon monoxide
EWL: C'E#O'Q



pyridine *N*-oxide
EWL: (C+N1'Q+C+C+C+C+), N1–O'E

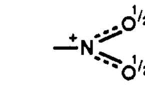
Note: For pyridine *N*-oxide, the N–O bond could not be represented as a dangler bond attached to N'Q because O also has 'E attached to it and +N'Q'O'E+ would be interpreted as being all attached to N, according to the convention of dangler bonds.

The resonance structure or delocalization often requires that the atoms in the structure be assigned a fractional charge. For example, the electronic configuration of formate ion is best represented as shown below in which each oxygen atom carries one-half the negative charge and the dotted line signifies one-half a bond[15] (i.e., delocalization).



In EWL the half-unit charge is represented by e (for negative) or q (for positive) attached by the dangler bond symbol to the appropriate atom. Thus, the formate ion is represented in EWL as O'e~C~O'e.

Sometimes, the coordination bond is associated with delocalization, e.g., the N–O bond in a nitro group. The EWL representation of the same is shown as



EWL: –N1'Q~O'e, N1~O'e

A number of full and half charges may be attached to the same atom, to represent one and one-half or two units of charge.

REPRESENTATION OF AROMATICITY IN EWL

A linear representation scheme is possibly tested to its limits while representing aromatic compounds. This is because for this class of compounds the chemical reality is the farthest removed from the atom-bond topological idealization. Further, the phenomenon of aromaticity may be defined in more than one manner,[7] being difficult to reconcile in some polycyclic and heterocyclic cases.

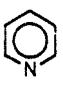
The present notation scheme is based on the following set of conjectures:

(i) A ring declared aromatic must have a conjugate cloud of electrons satisfying Hückel's condition ($4n + 2\pi$).

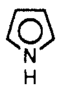
(ii) An atom that is a member of the ring and contributes one or more electrons to the π-electron cloud would be called an aromatic atom. Any bond connecting the aromatic atoms in a ring will be defined as an aromatic bond.

(iii) A ring atom that does not contribute to the electron cloud (e.g., the carbonyl carbon in tropone) would be called a nonaromatic member of the ring. Bonds connecting such atoms to neighboring ring atoms would be normal single covalent bonds.

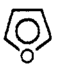
(iv) An aromatic atom contributing one electron to an aromatic ring would be considered normal, while one contributing two valence electrons (e.g., in cyclopentadienyl anion) or a lone-pair of electrons (e.g., in pyrrole, furan) would be considered as supranormal. The supranormal state is indicated by using lower case atom symbols for housekeeping of electrons and hydrogen atoms, e.g.



EWL: (C+C+C+N+C+C+)



EWL: (C+C+n+C+C+)



EWL: (C+C+o+C+C+)

Note that with the above assumptions:

(a) Two aromatic atoms in different rings can be connected by a nonaromatic bond, e.g., biphenyl (in Table I).

(b) Only one kind of aromatic bond need to be specified. The C–N bond of pyrrole and pyridine need not be differentiated by bond symbols. The difference would be apparent from the nitrogen symbol.

COMPARING EWL WITH THE ALTERNATIVES

In EWL, the use of dangler bonds reduces segmentation considerably while improving readability, e.g., mucic acid, hexachlorobenzene, and glucosamine (in Table I). The present system, on the average, is thus more compact compared to WLS. Also the functional group identification is easier. For example a carboxylic acid would be identified in EWL notation as –C″O′O or –C″O–O, ketone as –C″O–, and secondary hydroxyl as –C′O–.

Compared to the earlier scheme, EWL can distinguish more varieties of bonds, and further processing is not required for calculation of lipophilicity and molar refractivity using additivity schemes.[10] Also, it is possible to represent ions using EWL which was impossible in WLS.

CONOL-II has been compared with WLS in the Appendix. The EWL scheme retains all the advantages of WLS over CONOL-II. The presence of provisions for different types of bonds in EWL allows complete hydrogen suppression without any loss of information.

For acyclic compounds the effort required in encoding for SMILES is comparable to that of EWL. However, the latter representation may prove to be more readable. The single and aromatic bond suppression allowed in SMILES economizes on the length of the string, but affects readability. Further, there is no provision for representing a half-unit of charge in SMILES. Again, the hydrogen atom has to be "unsuppressed" or the structure has to be represented in aliphatic form in SMILES (see Table I) for representing pyrrole-like structures.

For polycyclic compounds the effort required to convert the polycyclic graph into a spanning tree, as is required in SMILES, is substantial. No such manual preprocessing is required for EWL.

COMPUTER PROGRAMS DEVELOPED

This author and co-workers have developed computer programs (in IBM-PC) using Turbo BASIC (TM Borland Inc.) for entering, editing, and error (syntactic and semantic) checking of EWL formulas.

The program (which constitutes several chained executable modules and runs comfortably on 256K bytes configuration) points out the type of error detected and the exact location of such error and permits immediate editing in a very user-friendly manner.

Error-free formulas are accepted for further processing which includes:

(i) Computation of molecular formula and molecular weight

(ii) Generation of Cambridge connection table

(iii) Computation of connectivity indices[12,13]

(iv) Estimation of partition coefficient and molar refractivity,[10] by additivity method

(v) Computation of van der Waal's volume[14]

In the process, the syntactic and semantic consistency of this notation scheme has been informally verified.

The structure, usage, and computer-science aspect of this software will be reported separately along with the distribution version of the software. Meanwhile, the entry and edit modules can be made available to interested noncommercial users on request.

CONCLUSION

With the modification suggested, the Wilcox and Levinson original linear notation scheme can be made more expressive, easier to learn, and applicable to a wider domain. This modified scheme is comparable to and in many respects even surpasses the capability of SMILES and CONOL-II. Advantages of this notation scheme may be utilized with the help of the computer programs developed. Immediate application of this methodology is expected to be in the area of quantitative structure–activity relationship studies.

ACKNOWLEDGMENT

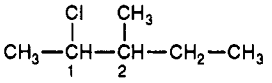
I thank Dr. S. C. Pakrashi, Ex-Director of the Institute, for his continued support, encouragement, and for providing facilities for this work and Dr. B. Achari for his continued encouragement and help through suggestions and criticisms. I am grateful to Dr. G. Chatterjee, Organon India, for critical evaluation of the methodology and the computer software. The software has been implemented with the help of Dr. T. K. Ghoshal, Professor, E. E., Jadavpur University, Calcutta. I have been supported by a fellowship from the Council of Scientific and Industrial Research, Government of India.

APPENDIX

A. Wilcox and Levinson Encoding Scheme. In WLS the atom symbols take their usual forms, e.g., carbon = C, nitrogen = N, chlorine = Cl, etc. Hydrogen atoms are left out. Single, double, triple, and aromatic bonds are denoted respectively by -, =, #, and + symbols. The delocalized double bond is also represented by the + symbol.

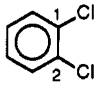
Straight chain compounds are simple to represent, e.g., CH_3-CH_2-Cl as C-C-Cl and CH_2=CH-Cl as C=C-Cl.

Branched chains require that each branch is represented by a separate segment and that the branching nodes be uniquely numbered. The segments are separated by commas. Node numbers are inserted after the corresponding atom symbols and the nodal atoms appear in each branch, e.g.



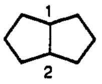
WLS: C–C1–C2–C–C, C1–Cl, C2–C

Rings are enclosed in parentheses and the last atom within the parentheses should be succeeded by the bond to the first atom of the ring. Cyclic segments, like linear segments, are also separated by commas.



o-dichlorobenzene
WLS: (C+C1+C2+C+C+C+), C1–Cl, C2–Cl

Fused rings should be considered as one or more rings plus a number of shunting/connecting branches, e.g.



[3,3,0]bicyclooctane
WLS: (C–C–C–C1–C–C–C–C2–), C1–C2
or (C1–C2–C–C–C–), C1–C–C–C–C2

B. CONOL-II. CONOL-II,[6] a linear notation developed by Hippe and co-workers, is being used in "all computer programs of higher order" in Poland. Basically it is akin to the WLS system except that the methodology of representation differs from WLS in the following respects.

1. Rings are not lexically specified. In WLS at least some rings are immediately identifiable by the parentheses.
2. Subsequent reference to a numbered node is by node number only, the atom symbol being dropped. This hampers readability.
3. CONOL-II counts the number of hydrogen with every non-hydrogen atom. This technique is helpful to establish Lewis structures with a limited number of bond symbols. Differentiating pyrrole-like bonds from pyridine-like bonds in the case of nitrogen is simple.[6] However, carrying this overhead for simpler compounds which form the majority is an unnecessary burden.
4. Nonavailability of bond types necessitates that one of the resonance structures must be selected for aromatic compounds and for compounds with delocalized bonds.

REFERENCES AND NOTES

(1) Ghoshal, N.; Basu, P. N.; Achari, B.; Ghoshal, T. K. Computer application to organic chemistry - A guided tour. *J. Inst. Eng. (India), Part CI* **1988**, *68*, 34–44.

(2) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer assisted studies of chemical structure and biological functions*; John Wiley-Interscience Publications: New York, 1979.

(3) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer storage and retrieval of generic chemical structures in patents. 6. An interpreter program for the generic structure description Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66–71.

(4) Heller, S. R. The development and evolution of a chemical information system. In *Data processing in Chemistry*; Hippe, Z., Ed.; Elsevier Scientific Publishing Co.: Amsterdam, 1980; p 177.

(5) Smith, E. G., Ed.; *The Wisewesser Line-formula chemical notation*; McGraw-Hill: New York, 1968.

(6) Hippe, Z.; Achmatowicz, O., Jr.; Hippe, R. Some problems of computer-aided discovery of organic synthesis. In *Data processing in Chemistry*; Hippe, Z., Ed.; Elsevier Scientific Publishing Co.: Amsterdam, 1980; p 210.

(7) Weininger, D. SMILES, a chemical language and Information System. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(8) Wilcox, C. S.; Levinson, R. A. A self-organized knowledge base for recall, design and discovery in organic chemistry. In *Artificial intelligence applications in Chemistry*; Pierre, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; p 228.

(9) Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and planning. In *Computer Assisted Organic Synthesis*; Wipke, W. T.; Howe, W. J., Eds.; ACS Symposium Series 61; American Chemical Society: Washington, DC, 1977; p 106.

(10) Ghosh, A. K.; Crippen, G. M. Atomic physiochemical parameters for three-dimensional structure-directed quantitative structure activity relationships interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.

(11) Watson, D. G. Lecture notes, Regional Seminar on Data Storage, Retrieval and Dissemination in science with special reference to chemical and molecular biosciences Madras, 1988.

(12) Kier, L. B.; Hall, L. H. *Molecular connectivity & drug design*; Stevens, G. D., Ed.; Academic Press: New York, 1976; Chapter 3, p 40.

(13) Pal, D. K.; Sengupta, C.; De, A. U. A new topochemical descriptor (TAU) in molecular connectivity concept: Part I. Aliphatic compounds. *Indian J. Chem.* **1988**, *27B*, 734–739.

(14) Morighuchi, I.; Kanda, Y.; Komatsu, Y. Van der Waals volume and the related parameters for hydrophobicity in structure–activity studies. *Chem. Pharm. Bull.* **1976**, *24*, 1799–1806.

(15) Griffin, R. W. *Modern Organic Chemistry*; McGraw-Hill Book Company: New York, 1986; Chapter 1, p 20.