SEFLIN—Separate Feature Linear Notation System for Chemical Compounds

CHI-HSIUNG LIN

Department of Chemistry, National Cheng Kung University, Tainan, Taiwan, Republic of China

Received June 27, 1977

A new chemical notation system designed for man-machine interactive treatment of chemical phenomena is described. Some WLN symbols are retained, but the encoding follows a new concept of separate features that allows the skeleton, π bonds, substitution groups, and special structural characteristics of the compound to be expressed as separate entries. Rules for encoding and decoding the notation as well as rules for transforming it in reactions are given.

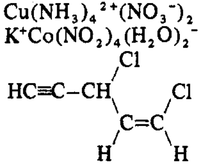
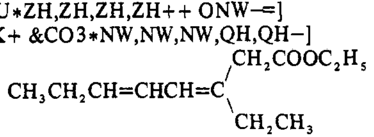
Generally, chemical linear notation systems emphasize the conserving of storage space in computer handling of chemical structures.[1,2] At the same time, to be specific for each individual compound, the notation should be unique and unambiguous. In this connection, Dyson's IUPAC notation system[3] and Wiswesser's linear notation (WLN) system[4] have been recognized as quite general and successful. As the number and types of compounds quickly increase, the conventional chemical nomenclature becomes more and more complex and difficult to handle by an ordinary chemist, and thus the nonconventional codified notation, originally designed for use on the computer, seems to be welcomed. The substitution of compound names by such systematic notations does not seem to be a matter of urgency as yet, but from the fact that chemists study their subjects mostly with structural formulas rather than with chemical names, the recent trend of studying a chemical system with methods generally known as artificial intelligence[5] requires a well-developed linear notation system as the database in the computer memory, which is equivalent to that of the structural formula to the chemist. Thus, a linear notation system should not be merely a device for computer indexing and retrieval of compounds but should possess the capability to be taken as the very object to be studied with methods familiar to chemists and more swiftly in the computer memory with modern mathematics in computer sciences.

In the past few years, studies have been conducted to develop algorithms to extract chemically significant features from a chemical notation in connection with such topics as automatic synthesis design[6,7] and pattern recognition.[8,9] The algorithm can be quite simple if there exists a notation system developed from the beginning for such purposes. A program which generates possible chemical structures on the knowledge of their fragments, such as DENDRAL,[10] will benefit from an efficient linear notation system. Such a system is also needed in the computerized study of chemical reactions.[11,12]

Under these circumstances, we have to choose either to improve an established notation system with more new symbolisms and rules so as to endow it with adequate flexibility or to develop an entirely new notation system as stated above. An example of the former choice is the ALWIN system introduced in 1974.[13] As for the latter choice, a new notation system applicable for both inorganic and organic substances is introduced in this paper.

This notation system originated from a digital notation system specifically designed for the study of chemical compounds using a desk calculator which does not allow character manipulation.[14] In the original 1974 version, a chemical structure was expressed with five groups of 12-digit numerals, each corresponding to the structure of sections of the skeleton, special atoms, the intrasection and intersection linkages, π-bondings, the substitution groups, and stereochemical descriptors in numerical codes. This digital notation could be transformed into a unique set of 12-digit numerals for the given substance by a calculator program. This system has been applied in various chemical studies in this laboratory in our effort to develop a new automated scheme of chemical inference on the behavior of simple organic compounds. During this period, we have accumulated experience and found the adequacy of the methodology. A new notation system which is so complete as to be capable of beng handled by an ordinary computer and used to generate unique and unambiguous linear notations of most chemical compounds is hence developed. In this new notation, four features of each compound are separately specified and put into a train of symbols. The first feature is the skeleton of the molecule expressed in a linear arrangement of symbols somewhat resembling WLN without the multiple-bond symbols and those for functional groups. The second and the third features are specifications of multiple bonds in the skeleton and functional groups attached, respectively. Stereochemical descriptors and other special terminologies are put in the last feature. Only alphanumeric characters acceptable by an ordinary computer are used for the symbol.

Table I. SEFLIN of Some Inorganic Compounds and Simple Open-Chain Organic Compounds

H_2O	[HQ]	CO_2	[OCO]	H_2SO_3	[QSOQ]	H_2SO_4	[QSWQ]
$Cu(NH_3)_4{}^{2+}(NO_3{}^-)_2$	[$CU*ZH,ZH,ZH,ZH++ ONW-=]						
$K^+Co(NO_2)_4(H_2O)_2{}^-$	[$.K+ &CO3*NW,NW,NW,QH,QH-]						
	[5/.1-2.1-2.4-5/3G5G/4CIS5]		[6<1VO2,2/.3-6]				

The separation of structural characteristics into four features wins for this notation system some facilities of fragment code notation systems such as the GREMAS code[15] and the IFI fragmentation code.[16] It simplifies the computer treatment of chemical characteristics to a considerable degree. What is more, the definition of a symbol in a given feature becomes clearer and easily manipulated in the computer. In this system, the hierarchy in encoding the structure is based mostly on groups in the skeleton only, so the change in functional groups seldom alters the symbol arrangement in the skeleton and hence the atom-numbering scheme which is implicitly embedded in the notation system. The "terminal ring first" strategy is adopted in encoding molecules with rings, which lessens the frequency of drastic changes of notation in the treatment of chemical reactions. A simple and natural sequence is adopted in encoding those complicated fused ring systems, for which no special symbol to express modes of the ring arrangement is necessary.

The planning of the present notation system makes chemically significant features more obvious to the computer[17] than in any other linear notation system currently in use. The

0095-2338/78/1618-0041$01.00/0

Table II. Symbols of Heteroatom Linkages in Organic Compounds

-CO- [V]	-S- [S]	-O- [O]	-N≡C- [N']
-COO- [VO]	-SO$_2$- [SW]	-NH- [N]	-C≡N- ['N]
-COOO- [VOO]	-SO$_2$NHSO$_2$- [SWNSW]	-N< [N<]	-N= [M]
-CONH- [VN]	-SO- [SO]	| -N- [N*] |	-NN- [MM]
-CON< [VN<]	-P(O)O- [P]		O ‖ -N- [L]
-CS- [U]	O ‖ -P-O- [PQ] | OH	-NHNH- [NN]	

Table III. Carbon-Containing Functional Groups To Be Included in the Skeleton of Organic Compounds (Forbidden for Use in Inorganic Compounds)

-C≡N [Y]	-COI [VI]	-COF [VF]
-CONH$_2$ [VZ]	-CH=O [VH]	-COBr [VE]
-COOH [VQ]	-COCl [VG]	-NHCN [NY]

Table IV. Special Symbols in the Skeleton

[+]	a positive charge
[<]	a two-way branching
[=]	a repetition of a group of symbols (at least three symbols)
['*n*']	*n*-membered carbon chain of any structure
['?']	any carbon chain
[#]	any skeletal group
[<(. . .)]	a bidentate ligand
[=%. . .%=]	a polymeric structure equivalent to $\{\ldots\}_n$
[−]	a negative charge
[*]	a three-way or higher branching
[@]	any aromatic group
[%. . . %]	a mutual linkage
[)]	chelation

assigned atom-numbering in the skeleton makes it convenient to point out any specific part or atom in the molecule. The importance of this capability in constructing a database of physical properties of compounds is emphasized by Skolnik in his linear notation system.[18,19] The natural scheme of atom-numbering of the present system facilitates its reference to the specific activity table which either exists as the database for the compound in the file or is generated automatically in accordance with some algorithm as the SEFLIN enters the computer.

The chemical change in reaction is manipulated in this system of notation very much like that in conventional structural formulas, and therefore it is treated rather easily in the environment of man-machine interaction. The structural change introduced as it is results in a notation which is not acceptable under the limitation of hierarchy rules. Simple mathematical rules are provided for this transformation to the approved notation in this paper for the convenience of chemists. The transformation proceeds automatically in the computer memory with a program to be introduced in the accompanying paper.[17] A special merit of the present notation system is that the conversion of the notation to the atom-environment table of the substance is extremely simple. The latter is a table of symbols expressing the connectivity of atoms and groups in the molecule and is utilized in the automated judgment on the possible reaction route, the automated generation of spectra, and the automated decision on physical, biological, or pharmacological properties by the computer.

This new linear notation system for chemical compounds is characterized by its special nature of separated structural features. It is, therefore, called the Separate Feature Linear Notation System or simply the SEFLIN system for chemical structures.

SYMBOLS AND RULES

The SEFLIN of a chemical substance is enclosed in a set of square brackets[20] and its four features are partitioned with slashes as follows:

[skeleton/π-bonding/substitution/terminology]

Symbols for atoms and groups are mostly similar to those adopted in WLN. Branch symbols [<] and [*] are used to suggest that the preceding atom is branched to link two or more parts in the skeleton, respectively. These linked parts are separated with a comma [,]. The whole molecule of inorganic compounds is deemed to be the skeleton. Elements other than H, C, N, O, S, and halogens are denoted with their atomic symbols preceded with a symbol [$] (read "string") or a symbol [&] (read "anded string"). The latter is followed, in addition, by a numeral expressing the oxidation number (see Table I).

Atoms in the skeleton are numbered implicitly in the sequence of the symbol arrangement. This numbering is used in the specification of the other features. The skeleton of organic substances includes (1) the carbon chain, (2) heteroatom linkages, (3) ring systems, and (4) special functional groups containing carbon atoms. The metal element in organic compounds is deemed to be a heteroatom linkage (see Tables II–IV).

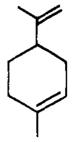
A multiple bond is encoded in regard to the π-orbital system. The symbol [.*n*] in the second feature denotes that the atom numbered *n* donates an atomic orbital to the π-orbital system. If every atom fron no. *n* to no. *m* does so, the notation is simplified into [.*n*-*m*]. Each atom with a triple bond donates two p orbitals to π_x and π_y orbitals, and is indicated twice with the symbolism.

When the molecule possesses a ring system, this is encoded first in the skeletal part of SEFLIN. This is the "ring-first" strategy in this notation system. An *n*-membered, isolated carbon ring is denoted as [R*n*)]. If it is a heterocycle, the heteroatom linkage is assigned in the fashion [R6.AO)] for tetrahydropyran. In this SEFLIN, the oxygen atom is assigned with the locant A. The locant assignment in the isolated ring is similar to that in WLN. All side chains on the ring are specified after the ring symbol with a leading ring locant. Some examples are shown in Table V.

A fused-ring system is denoted by assigning a terminal ring as its base ring and by successively forming new rings by ω,ω'-disubstitution of chains. These attached rings are called in sequence Ring-I, Ring-J, Ring-K, and so on. Locants on these arcs are assigned with the ring code IB, JC, etc. A four-membered arc on a six-membered ring constitutes the second ring in naphthalene or decalin. The skeleton is [R6)A4B)]. In this case, positions A and B on the base ring correspond to positions IA and IF on the second ring. Analogously, the skeleton of anthracene is [R6)A4B)IC4D)], where the arc from IC to ID is given the single preceding ring name I. The ring name is also omitted before the heterolinkage annotator before the symbol [)]. The assignment of the locant A in such fused-ring systems will be discussed later in the section of structural hierarchy. In Table VI, locants

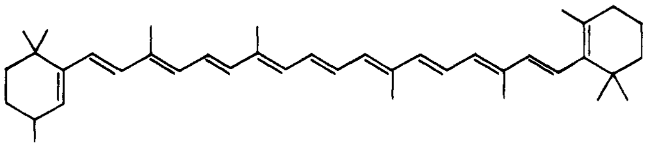
Table V. SEFLIN of Compounds with Isolated Rings

limonene



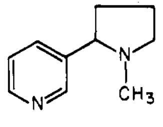
[R6)D1,A1<1,1/.3-4.8-9]

β-carotene



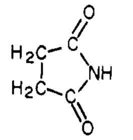
[R6)F1,B1,B1,A3<1,4<1,5<1,4<1,2R6)F1,B1,B1/ .1.6.1Ø-12.14-17.19-23.25-28.3Ø-32.37]

nicotine



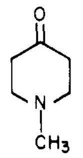
[R5.AN)A1,BR6.CM)/.7-12]

succinimide



[R5.AVBNCV)]

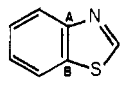
1-methyl-4-piperidone



[R6.AVDN)D1]

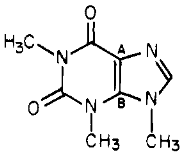
Table VI. SEFLIN of Compounds with Fused-Ring Systems

benzothiazole



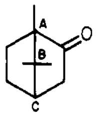
[R5.CSEM)A4B/.1-9]

caffeine



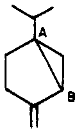
[R5.CNEM)C1,A4B.BVCNDV EN)IC1,IE1/.1–5.7–1Ø]

camphor



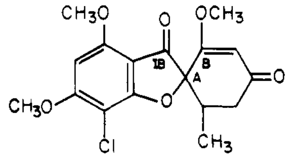
[R5.EV)A1,B1,B1,A2C)]

sabinene



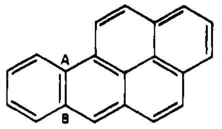
[R3)A1<1,1,A3B) ID1/.9–1Ø]

griseofulvin



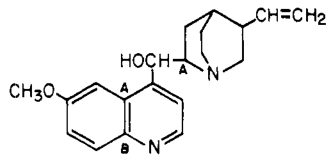
[R6.DV)F1,BO1,A4A.BVEO)IC4D) JBO1,JDO1/.2–4.1Ø–17/17G]

benzo[a]pyrene



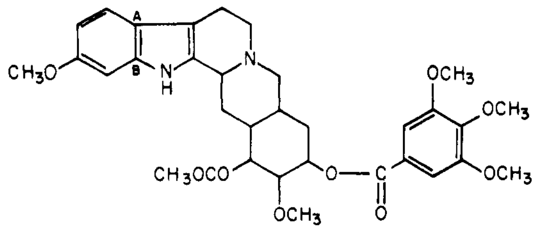
[R6)A4B)IC4D)IB3JB)JC3KD)/.1–2Ø]

quinine



[R6.CM)A4B)ICO1,F1R6.BN)D2, B2E)/.1–11.2Ø–21/13Q]

reserpine



[R6)DO1,A3B.DN)IB4C.DN)JD4E) KC4D)LDO1,LEOV1,LCOVR6)CO1, DO1,EO1/.1–7.9–11.3Ø–36.38.4Ø]

A and B are shown for the sake of clarity. The base ring is the smaller terminal ring.

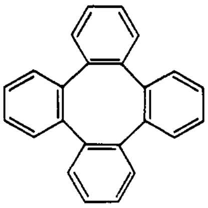
The above "ring + arc" strategy in the present notation system renders a logical atom numbering. A similar strategy is adopted in Skolnik's notation system[18,19] which emphasizes atom-based cognizance of molecular structure. The symbol [R] in the present system represents a ring system, either an isolated single ring or a system with fused rings. The number of the "chelation" symbol [)] gives a count of elementary rings under chemical sense.

The "chelation" symbol is used with either the ring or arc symbolism or the multiple branch symbol on a metal atom. The structural effect is the formation of a ring. This symbol may be preceded by a locant or a branch symbol. In such a case, it denotes that the corresponding atom is linked to the central metal atom. Another symbol which suggests a ring structure is a pair of "link" symbols, i.e., [%...%]. Examples are shown in Table VII.

A polymer with a definite structural unit is encoded with the "repeat link" symbol [=%...%=]. Polymers in general are expressed with the SEFLIN [=%#%=]. In the same manner, alcohols, phenols, ketones, aldehydes, carboxylic acids, etc., are expressed as [#//Q], [@//Q], [#V#], [#VH], [#VQ], respectively. No location is assigned in the third

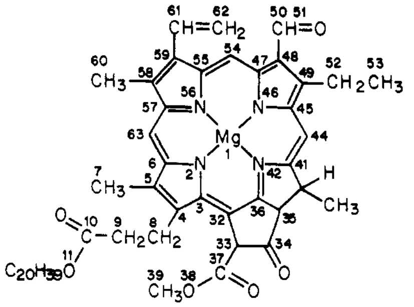
Table VII. SEFLIN with "Link", "Chelation", "Repetition", and "Bidentate" Symbols

tetraphenylene



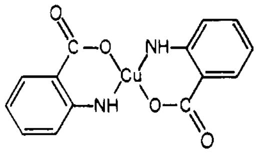
[R6)A%,BR6)BR6)BR6)B%/.1-24]

chlorophyll *b*



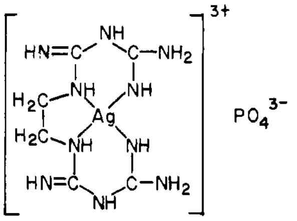
[$MG*R5.AN)E%,D1C2VO'20',B1R5.CV)
BVO1,D3E.DM)D),B1,C1R5.BN)B),DVH,
E2,C1R5.BM)B)D1,E2,C1%/.2-6.32.36
.40-42.44-50.54-59.61-63//'20'39H]

copper anthranilic acid



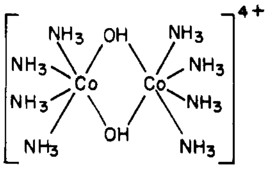
[$CU*<(OVR6)BN)=/.2-10,=]

silver ethylenedibiguanidium phosphate



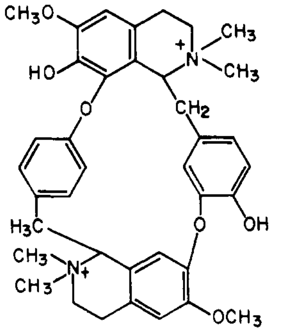
[&AG3*N1N1N<),2N<),1N1N)+++OPWO
---/.2-6.9-13/2Z5M10M12Z]

octaammine-*μ*-dihydroxodicobalt(III)



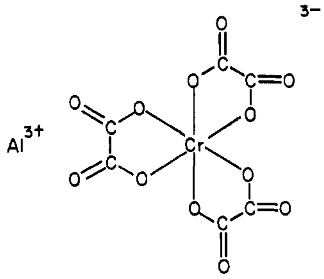
[&CO3*ZH,ZH,ZH,ZH,Q%,
Q&CO3*%,ZH,ZH,ZH,ZH++++]

tubocurarine



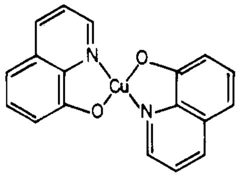
[R6)CO%,EO1,A4B.DN)ID1,ID1,IE1R6)
COR6)BO1,D4E.DN)ID1,IF1R6)D%++
G-G-/.1-7.9.15-18.27-32/4Q17Q]

aluminum trioxalatochromate



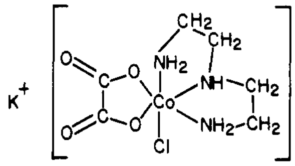
[$AL+++&CR3*<(OVVO)==---]

copper oxime



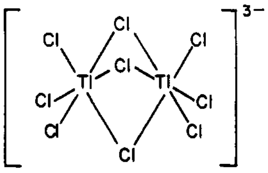
[$CU*<(OR6)B4C.BM)IB)=/.2-12,=]

potassium chlorooxalato(diethylenetriamine)cobaltate



[$.K+ $CO*G,<(OVVO),N2N<),2N)-]

nonachlorodithallate ion

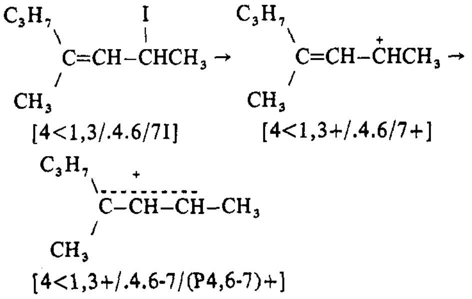


[$TL*G,G,G,G%,G%%,G$TL*%,%%,G,G,G---]

feature of alcohols and phenols. This suggests that any location in [#] and in [@] is possible (see Tables VIII and IX).

When a substitution group is located in a certain region of the molecule, the region is specified in parentheses and put in the third feature of SEFLIN. Bromo-α-naphthol with the bromine atom located on the phenol ring, for instance, is encoded thus: [R6)A4B)/.1-10/3Q(4-6)E]. A question mark (?) is used if an uncertain group is in the skeleton, e.g., ethyl-α-naphthol [R6)?2,A4B)/.1-10/3Q].

In the third feature of SEFLIN, symbols [+], [-], and [*] suggest a filled, a vacant, and a half-filled orbital which may substitute a hydrogen atom at the given location. A π orbital with negative charge, positive charge, or spin density distribution is also encoded with these symbols; the location is assigned with parentheses as before and preceded with a character P.



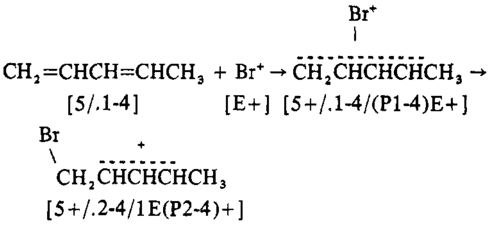
[4<1,3/.4.6/7I] [4<1,3+/.4.6/7+]

[4<1,3+/.4.6-7/(P4,6-7)+]

Analogously, benzene anion radical produced for ESR study is encoded thus:

[R6)–/.1-6/(P1-6)*]

The complexing of a cation on the π-orbital system results in the formation of a π complex.

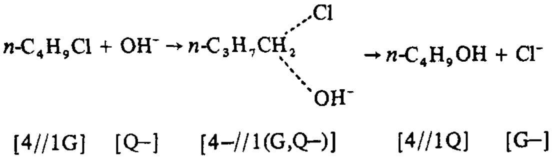


[5/.1-4] [E+] [5+/.1-4/(P1-4)E+]

[5+/.2-4/1E(P2-4)+]

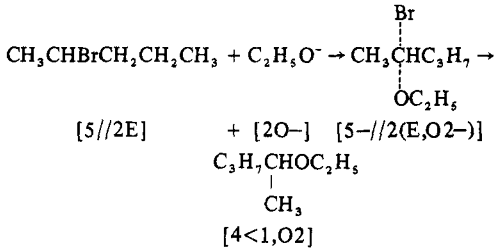
The complexed cation is put into the third feature of SEFLIN even when it is a carbonium ion with a complicated structure. An example is aniline complexed with $F_2CHCH_2C^+(CH_3)_2$; the SEFLIN is [R6)+/.1-6/1Z(P1-6)1*1,1,2<F,F+].

The transition state of an S_N2 reaction or an aromatic substitution reaction is often represented with a σ complex, in which a hydrogen atom is replaced by two groups. Thus,



[4//1G] [Q–] [4–//1(G,Q–)] [4//1Q] [G–]

Here, the reaction corresponds to the interchange of substitution groups. In other cases, the incoming group becomes a part of the skeleton, e.g.



[5//2E] + [2O–] [5–//2(E,O2–)]

[4<1,O2]

The transition state of an S_N1 reaction can be considered in the same way. Then the reaction center is a σ complex with the leaving group and a vacant atomic orbital under formation. The S_N1 transition state for 2-bromopentane becomes [5//2(E,+)]. The transition state of a radical reaction may appear as follows:

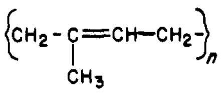
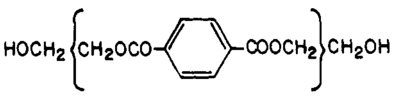
$CH_3CH_3 + Cl\cdot \rightarrow CH_3CH_2\text{---}HCl \rightarrow CH_3CH_2\cdot + HCl$

[2] [G] [2//1(HG,*)] [2//1*] [HG]

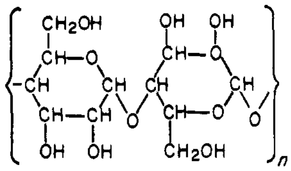
STRUCTURAL HIERARCHY

The structural hierarchy in this notation system decides which part of the molecule should be encoded before other parts. The resultant SEFLIN is quite similar in its symbol order to the conventional linear formula of chemical structure.

Table VIII. SEFLIN of Polymers

rubber	Dacron
	
[=%2<1,2%=/.2.4]	[1=%1OVR6)DVO1%=1/.3-12/1Q14Q]

cellulose



[=%R6.CO)B1,DOR6.CO)B1,DO%=//5Q6Q7Q13Q14Q15Q]

The first part to be encoded is the most complicated ring system or, if there is no ring in the molecule, the most complicated chain. If the substance is ionic, then the cation is encoded first.

The complexity of a ring system is decided by the number of elementary rings it contains. If the number is the same, then higher hierarchy is bestowed on one with smaller rings. Heterogeneity, unsaturation, and side chains are looked for, if the decision is difficult.

The complexity of a chain is decided by looking at (a) the number of SEFLIN symbols, (b) the total number of atoms, (c) the number of atoms in its main chain, and (d) the first numeral as a SEFLIN symbol. The hierarchy sequence of alphabetic characters in the skeleton of SEFLIN is Y–V–U–T–S–P–O–N–M–L–#–@–'–&–$–%. The first symbol of SEFLIN of organic compounds should be [R] or a numeral. Symbols [Y] and [V] (for –C≡N and for >C=O) are allowed only when there is no terminal ring or chain, e.g., formaldehyde [VH], oxalic acid [VQVQ], urea [V//ZZ].

Groups following a branch symbol or a ring symbolism are arranged from a lower to a higher hierarchy, e.g., n-$C_4H_9N(CH_3)(C_2H_5)C_3H_7^+\ I^-$ [4N*1,2,3+I–]. Atom numbering is given in the natural sequence in the skeletal part of SEFLIN. Thus the atom numbering for the nitrogen atom in this example is 5, and the iodide ion 12. A number is assigned for every alphabetic character in the heteroatom linkage symbols (Table II) except [W] which is not numbered; n numbers are given for the carbon chain expressed with the general symbolism ['n']. Atoms in the "any group" symbols ['?'], [@], and [#] are not numbered.

The first ring system is encoded from a special base ring. This is a terminal ring in this system. In a multilayered fused-ring system, the base ring is the terminal ring of the longest possible first layer, so as to make the successive encoding of rings proceed layer-wise in the same direction. The base layer is assigned in view of complexity if length is not applicable. The atom numbering starts from one of the fusion sites on the base ring. This is also the location A. In the multilayered system, it is the "upper" fusion site. In a single-layered fused-ring system, the hierarchy for assigning the base ring is judged by the ring size, heterogeneity, unsaturation, and side chains. After the location A is assigned, the adjacent position which is nearer to the other fusion site is given a locant B and an atom numbering 2.

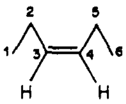
The other ring systems are encoded from the ring linked to the first ring system. The position of the linkage is the location A, and the numbering continues from this atom.

Table IX. Symbols for Ordinary Substitution Groups

-NH$_2$ [Z]	-OH [Q]	-NHNH$_2$ [NZ]	=NOH [MQ]
-halogen [X]	-PO(OH)$_2$ [P]	-NO$_2$ [NW]	=NH [M]
-SH [T]	-ONH$_2$ [OZ]	-NHOH [NQ]	-I [I]
-SO$_2$NH$_2$ [SWZ]	-OSO$_2$OH [OSWQ]	-NO [NO]	-Cl [G]
-SO$_2$OH [SWQ]	-OOH [OQ]	-N≡C [NC]	-F [F]
-SOOH [SOQ]	-OPO(OH)$_2$ [OP]	=NNH$_2$ [MZ]	-Br [E]
-PO(OH)OPO(OH)OPO(OH)$_2$ [PPP]		-OPO(OH)OPO(OH)OPO(OH)$_2$ [OPPP]	
-PO(OH)OPO(OH)$_2$ [PP]		-OPO(OH)OPO(OH)$_2$ [OPP]	-N$_3$ [A]

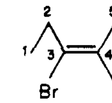
Table X. Numbering Schemes for Stereochemical Descriptors

cis-3-hexene



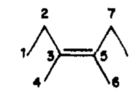
[6/.3–4//3CIS4]

cis-3,4-dibromohexene-3



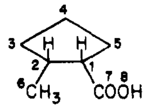
[6/.3–4/3E4E/3CIS4]

cis-3,4-dimethylhexene-3



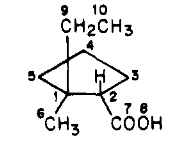
[3<1,1<1,2/.3.5//4CIS6]

cis-3-methylcyclopentanecarboxylic acid



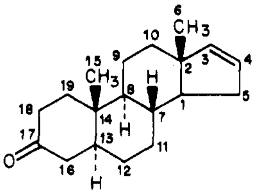
[R5)B1,AVQ///6CIS7]

3-methyl-*trans*-3-ethyl cyclopentanecarboxylic acid



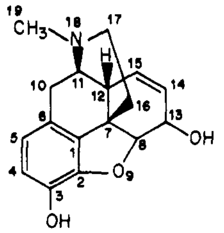
[R5)A1,BVQ,A2///7TRANS9]

androstenone



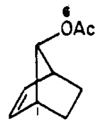
[R5)B1,A4B)IB4C)JE1, JD4E.CV)/.3–4//2A, 7A,8B,13B,14A]

morphine



[R6)A3B.DO)F3IB)IC3JD) IB3JC.DN)LD1/.1–6.14 –15//7A,11A,12A]

anti-7-norbornenyl acetate

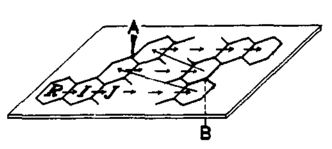


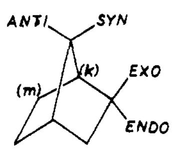
[R5)BOV1,A2C) .4–5//6ANTI(1)]

STEREOCHEMICAL DESCRIPTORS

Stereochemical descriptors are assigned in the fourth feature of SEFLIN. To let the computer accept the descriptor, rules are set up for a more specific encoding. Examples are shown in Table X.

Descriptors A and B correspond to upside and downside bonds on the plane of the fused ring system illustrated below. Reference atoms *m* and *k* in the descriptor notations *n*ENDO(*m*), *n*EXO(*m*), *n*ANTI(*k*), and *n*SYN(*k*) are defined below.





TEMPORARY SEFLIN

In the automated judgment of reaction route by computer, SEFLIN is always transformed into the atom-environment table on which possible chemical changes, such as the bond formation, the bond cleavage, and the hydrogen transfer,[21] are manipulated. The table is then rearranged by a specific subroutine for output of the SEFLIN of the product. A simple example is the ring formation of farnesol [3<1,4<1,4<1,1/.2-3.7-8.12-13/1Q] on dehydration, which gives bisabolene [R6)D1,A1<1,4<1,1/.1.3-4.8.12-13], a typical biosynthesis. The transformation is actually simple for the computer but is not an easy matter for the chemist. A transformation rule such as ([%*n*<%,*p*]) → [R*n*)A*p*] could be helpful. The former notation is a temporary SEFLIN (t-SEFLIN) for the consideration. Other examples of t-SEFLIN are:

([*p*//*n*#]) ↔ ([*n*<#,*m*]), where *p* = *n* ⊕ *m*,
 e.g. [3//2Q] → ([2<Q,1])
[*p*O#] → ([*p*//*p*(O#)]), etc., e.g. [3O2] → ([3//3(O2)])
[*p*V#] → ([*q*//*p*+,*p*O−]), where *q* = *p* ⊕ 1 ⊕ #,
 e.g. [3V2] = ([6//4+,4O−])

A t-SEFLIN may be one generated in the process of chemical transformation. Rearrangement of its skeleton in accordance with the hierarchy rules changes it to a good SEFLIN. To make this rearrangement, the t-SEFLIN is first partitioned into units of symbols; each of these units is a part in the main chain structure. For example,

[R6)D2,A4<2,4*1,3,OV2] → ([R6)D2,|A3|1<2,|3|1*1,3,|O|V|2])

Interchange of groups inside a unit is allowed without modification, i.e., ([a|1*p,q,|z]) → ([a|1*q,p,|z]), etc. A group in a unit can be exchanged with the whole part following this unit, i.e., ([a|1*p,q,|z]) ([a|1*p,z,|q]), etc. Exchange of units is forbidden. The inversion of the whole t-SEFLIN is accomplished by reversing the order of all units and inverting both the first and the last units, i.e., ([a|b|...|y|z]) → ([z'|y|...|b|a']), where a' and z' are inverted a and z, respectively. A group in a unit can be exchanged with the whole part of the t-SEFLIN; both are inverted, i.e., ([a|b...h|1*p,q,|z]) → ([q'|1*p,h...ba',|z]). These and other transformation rules are applied when considering chemical behavior of substances with SEFLIN.

ACKNOWLEDGMENT

The author is grateful to Professor Y. Yoneda of the University of Tokyo, Japan, for valuable discussions at the initial stage of the formulation of this notation system. He thanks Mr. Hwa-Fu Chen for his assistance during the preparation of the manuscript.

REFERENCES AND NOTES

(1) C. H. Davis and J. E. Rush, "Information Retrieval and Documentation in Chemistry", Greenwood, Westport, Conn., 1974.
(2) J. E. Rush, "Status of Notation and Topological Systems and Potential Future Trends", *J. Chem. Inf. Comput. Sci.*, **16**, 202 (1976).
(3) IUPAC, "Rules for IUPAC Notation for Organic Compounds", Commission on Codification, Ciphering and Punched Card Techniques of the IUPAC, Wiley, New York, N.Y., and Longmans, Green, London, 1961.
(4) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation",

McGraw-Hill, New York, N.Y., 1968.

(5) More than 20 papers under the main title "Application of Artificial Intelligence for Chemical Inference" by the Artificial Intelligence Group of Stanford University have appeared in *J. Am. Chem. Soc.* since 1969. Pattern recognition is a branch of artificial intelligence familiar to chemists. Papers of general interest can be found in Computer and Control Abstracts, Section 6.440.

(6) E. J. Corey, W. T. Wipke, R. D. Cramer III, and S. J. Howe, "Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecule", *J. Am. Chem. Soc.*, **94**, 431 (1972).

(7) J. B. Hendrickson, "Systematic Synthesis Design. IV. Numerical Codification of Construction Reactions", *J. Am. Chem. Soc.*, **97**, 5784 (1975).

(8) A. J. Stuper and P. C. Jurs, "ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Technique", *J. Chem. Inf. Comput. Sci.*, **16**, 99 (1976).

(9) W. E. Brugger, A. J. Stuper, and P. C. Jurs, "Generation of Descriptors from Molecular Structures", *J. Chem. Inf. Comput. Sci.*, **16**, 105 (1976).

(10) D. H. Smith, L. M. Masintes, and N. S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structure", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley, New York, N.Y., 1974.

(11) M. Osinga and A. A. V. Stuart, "Documentation of Chemical Reactions. III. Encoding of the Facets", *J. Chem. Inf. Comput. Sci.*, **16**, 165 (1976).

(12) G. W. Adamson and D. Bawden, "An Empirical Method of Structure-Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **16**, 161 (1976).

(13) E. V. Krishnamurthy, P. V. Sankar, and S. Krishnan, "ALWIN-Algorithmic Wiswesser Notation System for Organic Compounds", *J. Chem. Doc.*, **14**, 130 (1974).

(14) C-H. Lin, "Computer on the Lab Bench—A Primary Step toward Computer Chemistry", *Kagaku no Ryoiki*, **28**, 771 (1974).

(15) S. Rossler and A. Kolb, "The GREMAS System, an Integral Part of the IDC System for Chemical Documentation", *J. Chem. Doc.*, **10**, 128 (1970).

(16) M. Z. Balent and J. M. Emberger, "A Unique Chemical Fragmentation System for Indexing Patent Literature", *J. Chem. Inf. Comput. Sci.*, **15**, 100 (1975).

(17) C-H. Lin, "Chemical Inference Based on SEFLIN. I. Basic Cognizance of Molecular Shape, Fragments, and Atomic Environment of Organic Compounds", *J. Chem. Inf. Comput. Sci.*, following paper in this issue.

(18) H. Skolnik, "A New Linear Notation System Based on Combinations of Carbon and Hydrogen", *J. Heterocycl. Chem.*, **6**, 689 (1969).

(19) H. Skolnik, "A Computerized System for Storing, Retrieving, and Correlating NMR Data", *Appl. Spectrosc.*, **26**, 173 (1972).

(20) The set of square brackets is an identifier and is not included in the computer memory as part of SEFLIN.

(21) Hydrogen atoms in the molecule become apparent as the SEFLIN is transformed into the atom-environment table; see ref 17.

Chemical Inference Based on SEFLIN. I. Basic Cognizance of Molecular Shape, Fragments, and Atomic Environment of Organic Compounds

CHI-HSIUNG LIN

Department of Chemistry, National Cheng Kung University, Tainan, Taiwan, Republic of China

Received September 6, 1977

Algorithms are presented to obtain basic information in the computer memory from input SEFLIN of organic structures. The group connectivity table and the atom environment table with a ring locant table are used as bases for treating physical and chemical behavior of organic substances. A scheme of automatic generation of descriptor sets from SEFLIN for pattern recognition is shown. A subroutine for rearranging the atom environment table after the chemical transformation of the compound in reaction is presented.

The Separate Feature Linear Notation (SEFLIN) is a new structural formula system for chemical study under the man–machine interactive environment.[1] It is chemist-oriented in that the encoding is rather straightforward from the conventional structural formula. It is also machine-oriented because it is easily generated in the computer if some chemical characteristics are assigned. To a computer, SEFLIN is a communication device with the chemist as well as a database with which chemical behavior of the corresponding substance is to be studied.

The chemist conceives the chemical nature of a substance from the structural formula, and its fragments and their combinational characteristics. These are equally evident from the SEFLIN, and therefore the chemist may learn to look into the SEFLIN through the chemical common sense he already possesses. This common sense is difficult for a computer to acquire. In this series of papers, programming strategies which endow the computer with such chemical common sense will be demonstrated.

The term "chemical inference" is quite vague. It may mean any chemically significant judgment based on molecular structure which is thought to be approved by most chemists. It may also include schemes of deduction most chemists prefer in reaching an assumption or conclusion. It may be strictly logical or somewhat intuitive. But never is it absolutely accurate. This is because the object of chemistry is a huge number of substances under infinitely diverse situations. A valid chemical inference has a statistical nature. Thus, a better inference may be what leads to a prediction closer to the statistical means of those tested substances. Since the number of tested substances is continuously increasing, any approved inference is always subject to improvement. What is really important, therefore, is to settle down to exploit the algorithm with which a certain chemical inference can be made by the average chemist or, as we hope, the computer.

There are several levels of structural problems on which an inference is made. We may look at the size and shape of the molecule. We may say there does exist a certain functional group. We may find that a certain atom or group is situated at a certain position in reference to another atom or group. We ask for the electron density distribution. We are even interested in such matters as the through-the-space interaction of electronic or nuclear spins. All of these are studied by the chemist on the same structural formula of the substance. This is to say that the conventional structural formula is successful in providing levels of information. But, we should not forget the hidden factor which is the amazing human intelligence. We look into the structural formula with a proper depth of focus for the occasion and deduce therefrom different concepts for the specific purpose. This is the mysterious power of cognizance. The artificial intelligence we need to build for the study of chemistry should be endowed with the capability

0095-2338/78/1618-0047$01.00/0