CHEMO Notation. A Line Notation for Organic Compounds Following IUPAC Nomenclature

Yukio Yoneda[†]

Institute of Research and Development, Tokai University, 2-28-4, Tomigaya, Shibuya-ku, Tokyo 151, Japan

Received September 6, 1995[⊗]

CHEMO notation is a high-level line notation faithfully in line with the IUPAC nomenclature to describe organic compounds and even reactions. It has conserved the concepts in nomenclature concerned with superatoms (parent rings and functional groups) as well as stereochemistry, by describing the corresponding symbols and modifiers as universal descriptors that may be used in estimation in chemistry. It can describe fusion, atom replacement, configuration, etc. One can convert a CHEMO formula or equation both to a connection table composed of superatoms as nodes and stereo modifier symbols and to a conventional one composed of single atoms and vice versa. CHEMO notation describes compounds more precisely than prevailing line notations that denote only skeletal connection in them. It does not request rigorous rules, but a subsystem CANON can prepare unique representation from it.

INTRODUCTION

Historically,[1−3] the first methods of input of chemical structures were line notations such as IUPAC notation[4] or Wiswesser Line Notation,[5] both requesting severe restrictions as they aimed at the unique and unambiguous notation of complicated compounds. Both CHEMO notation[6,7] and SMILES notation,[8] by Weininger in 1988, are their successors, but aiming at no uniqueness in notation; preparation of the unique notation is left for computer processing. All notations except CHEMO notation, however, describe a compound essentially with atom symbols by decomposing it into the constituting atoms. The second method was that of topological connection tables, described and punched manually by users. The third method consisted of two-dimensional chemical diagrams, drawn manually on a display. Drawing chemical diagrams on a display, however, demands not the least skill and sweat in the case of complicated compounds. The fourth method was to use unambiguous chemical names such as STARS name.[9] A long name, however, may suffer from possible mistakes in typing; sophisticated programs with large databases may be indispensable for decoding it to a connection table. The fifth and newest method is recognition of chemical diagrams,[10,11] printed or drawn by hand; the tool for recognition has been proven as reliable.

The second, manually described connection table has become entirely obsolete. Among input methods mentioned above, contemporarily the most appropriate ones may be drawing on a display or reading chemical diagram with an image scanner. Compounds described with the first, and the third through fifth methods are finally converted to connection tables appropriate in the systems employed. Most of the converted tables, however, consist of an atom table storing single atoms as nodes and a bond table storing bonds among atoms. Some of them are accompanied by stereochemical modifiers but usually they are *relative* modifiers such as “parity”.[12]

Traditionally, chemists have shared as their common precious knowledge the concepts of superatoms, that is, parent rings (including locants) and functional groups as well as *absolute* stereochemistry, etc., all in chemical nomenclature. These concepts have suggested to chemists the static and dynamic nature of chemical compounds and therefore stimulated their imagination. For example, parent ring or functional group names let chemists foresee specific physical properties due to the neighboring effects of higher orders as well as global effects such as aromaticity.

In the prevailing systems estimating physical or biological properties from chemical structures, single atoms in a connection table are usually encoded to superatoms, or descriptors, but the descriptors are often arbitrarily defined in each system. IUPAC nomenclature[13] of organic compounds employs, as *universal* descriptors throughout organic chemistry, the selected names of parent rings and functional groups. Following IUPAC rules, indeed, a name, i.e., a linear string of descriptors, represents an unambiguous skeletal structure, functional groups, and even stereochemistry by inserting well-defined descriptors such as (*R*), (*Z*), or (*M*).

CHEMO notation has been used in CHEMOGRAM, shown below, as input and intermediate language in it. It utilizes *symbols*, derived from IUPAC names or rational formulas, as universal descriptors to represent superatoms and rigorous stereochemistry in connection tables. Some examples are as follows: BENZ for benzene ring, NHCONH for ureido, and .R. for (*R*). Moreover, it can describe molecules and even reactions in line notation with a grammar similar to IUPAC rules. CHEMO notation is convertible to a connection table composed of superatom as nodes and associated by stereo modifier symbols and vice versa. Because the line notation needs much less space as compared with a connection table to describe a given compound, CHEMO notation may serve as a convenient substitute for a connection table. It may be noted that we can easily understand chemical structure by reading CHEMO notation as shown later, in contrast to that we hardly do by reading a connection table.

As the first function of CHEMO notation, it defines a variety of universal descriptors, derived from IUPAC no-

[†] Current address: 3-1-37-2316, Hontamon, Tarumi-ku, Kobe 655, Japan. Tel/Fax +81-78-787-2740.

[⊗] Abstract published in *Advance ACS Abstracts,* January 15, 1996.

0095-2338/96/1636-0299$12.00/0

menclature, and they may be described in connection tables. Its second function is to describe chemical structures of organic compounds following IUPAC rules. Let us explain the detail of the second function. CHEMO notation performs a high-level line notation, or an artificial language, faithfully in line with the IUPAC nomenclature. Consequently, CHEMO notation is very similar to the conventional rational formula described with ring and functional group names. In principle, it bears close resemblance to SMILES as far as chain structures concern. For representing complicated ring structures, CHEMO notation allows even ring syntheses, such as fusion, spiro, and bridging among basic rings, and ring modifications such as atom replacement and hydro or dehydro reformation within rings. Stereochemistry is also strictly described by inserting similar modifiers in the string.

CHEMO notation, however, not at all requests such strict rules to describe unique formulas as does Wiswesser Line Notation; for retrieval and identification; the former will be made canonical, or unique, by a subsystem CANON prepared by us. Consequently, the described CHEMO formula is not a cipher but a modified rational formula; therefore chemists can easily understand the compound without such trouble as counting number of letters and delimiters. Connection tables derived from CHEMO notation may serve as input to any existing computer programs.

CHEMOGRAM[6,7] means an integrated computer program package for creation and estimation of chemical substances and reactions with the aid of chemical logic; in other words, it means a set of expert systems aiding chemists in designing them. Major systems in CHEMOGRAM consist of STERIC,[7,14] EMPRIC[6,7,15]/EROICA,[6,7,16] and GRACE.[17] All of them are composed of program packages and databases, details of which will be reported successively.

1. STERIC. It will prepare Cartesian coordinates of organic compounds.

2. EMPRIC/EROICA. EMPRIC will estimate, and EROICA will retrieve and estimate, by integration with EMPRIC, fundamental physical properties of organic compounds.

3. GRACE. It will create gaseous phase radical reactions and their Arrhenius parameters and then predict their reaction rates and selectivity among products.

In general, CHEMOGRAM programs employ connection tables described with CHEMO symbols as working tables and describe intermediate fragments, compounds, and even reaction formulas with CHEMO notation to save memory and to secure visible readability. These systems also permit CHEMO notation as input and as description of chemical structure in databases.

We first published the detail of CHEMO notation (first version) and the concept of CHEMOGRAM in Japanese[6] in 1970 and then published an outline in English[7] in 1975. Recently we have completed the fifth revised version of CHEMO notation, which will be presented here.

OUTLINES OF CHEMO NOTATION

CHEMO notation adopts the concept of *superatoms* (atom groups) and modifiers within chromatic chemical graphs. We designate as superatoms the *base groups* (parent rings or unsubstituted ring structures with locants) and *hetero groups* (functional groups) as well as single atoms with attached hydrogen(s). Popular base groups are cited by abbreviated symbols, and any other base groups may be described by reformation and synthesis of base group symbols. Ring reformation consists of atom replacement, hydro/dehydro description, and homologation of rings (ring enlargement) like homoreformation in steroid. Ring synthesis consists of fusion, bridge, spiro and "join" that unifies two rings to make a large ring. These descriptions are realized by writing corresponding modifiers. A hetero group is described by a string of element symbols and multiple bonds; they have been chosen from major functional groups conventionally used in IUPAC nomenclature.

Modifier symbols written by alphanumerals between two periods, like logical operators in FORTRAN, are employed to describe almost all specification, including stereochemistry, in nomenclature. Major examples of specification are locant, atom replacement, ion, isotope, configuration such as cis/trans, (*R*)/(*S*), (*Z*)/(*E*), and (*P*)/(*M*), conformation, and reformations, syntheses and type of rings, and description of numerical values in a three-dimensional structure. Thus, in principle, any organic molecule can be described with superatom descriptors, modified by modifier symbols, combined by bond symbols, and nested by parentheses.

SPECIFICATION AND RULES

CHEMO notation denotes a molecular structure as a sequence of literal elements that is essential in chemical names following chemical nomenclature. Only prerequisite is proper choice of symbols for superatoms, i.e., base group(s) and hetero group(s). CHEMO notation allows to place any space in the string of characters. It does not claim a canonical or unique representation; therefore, there are usually a few ways of valid description for a structure.

Elements of description in CHEMO notation consist of substance symbols, modifier symbols, bond symbols, and delimiters. All symbols in CHEMO notation are described by capital letters, except for atom symbols which are written by a capital letter followed by a lower-case letter for two-letter symbols.[18]

We describe CHEMO notation of a compound by connecting substance symbols by bond symbols[19] *without* omitting single bonds; branches as well as substituents to ring are specified by enclosure in parentheses. The following symbols represent bonds: − (single); ═ (double); −═ (triple); ,− (aromatic); − − (molecular bond); and ,, (ionic bond). Full sets of symbols appear in the Appendix, although its reasonable subsets may be allowed by option under specified environments.

1. Substance Symbols. Substance symbols consist of (1) atoms, (2) base groups, (3) hetero groups, and (4) inorganics.

(1) Atoms Symbols. (a) Ordinary Atoms. Atoms from hydrogen (H) to uranium (U) are represented by their element symbols; D may be used for deuterium in place of He by option, with a hypothetical atomic number of 2. Description of attached hydrogen(s) with an *atom count* (1 may be neglected) is compulsory if an atom is associated by hydrogen atom(s). Saturation or unsaturation of normal valency of an atom may be implicitly described by the number of attached hydrogen atoms. Except for attached hydrogen, the atom count is prohibited after atom symbols.

C	carbon atom (not methane)
CH4	methane
N	nitrogen atom (not ammonia) or amino radical (tertiary)
NH2	amino radical (primary)
Cl	chlorine atom (not hydrogen chloride)
Au	elemental gold
H−H	hydrogen molecule
CH3−CH−CH3	2-propyl radical (implicit expression)

(b) Pseudoatoms. Some symbols represent radicals, intramolecular linkage, empty atoms, and active sites over catalysts. These are described in CHEMO notation as if they were ordinary atoms.

R	radical (explicit expression)
CH3−CH(−R)−CH3	2-propyl radical
Xi, Zi	*intramolecular linkage* and ring closure
CH2(−X1)−CH2−CH2−CH2−CH2−CH2(−Z1)	cyclohexane (extended)
CH(−X1)=CH−CH=CH−CH=CH(−Z1)	benzene (extended)
E, E+, E−	*empty* atoms
−E+	hydro reformation of a ring
−E	dehydro reformation of a ring
S1′, S2′	active sites over heterogeneous catalysts

(2) Base Group Symbols. Base groups indicate methane and parent rings. CHEMO base group symbols, four-character or six-character alphanumeric abbreviations, represent selected unsubstituted rings that appear more frequently in compounds. Besides, any ring structure may be represented by a chain structure ring-closed with intramolecular linkage as shown above in case of cyclohexane and benzene. Acyclic or cyclic compounds are described as derivatives of base groups, methane, or appropriate parent ring(s), respectively.

About 600 base group symbols consists of ordinary and additional base groups; each of them is given a unique abbreviation that is alphanumeric where the beginning letter and all sequential letters are capital letters.

(a) Ordinary Base Groups. Four-Character Abbreviation. They are allowed throughout all CHEMOGRAM programs. In version 5, we selected around 244 ordinary base groups based on the following standpoints: (i) parent hydrocarbon rings (Rule A-21, 21.2) and parent heterocyclic rings (Rule B-2, 2.11) and trivial and semitrivial names which have been retained for compounds and as bases of fusion names in the IUPAC Nomenclature 1979,[13] (ii) popular monocyclic cycloalkanes and cycloalkenes whose number of ring members is 10 or smaller, (iii) a few bridged rings, and (iv) a variety of popular rings. Some examples are as follows:

BENZ	benzene (i)
CYHA	cyclohexane (ii)
C6H*n*	benzene ($n \leq 5$) or cyclohexane ($n \geq 7$) (iv)
DECL	decahydronaphthalene (decalin) (iv)
PRDN	pyridine (i)
NBNA	8,9,10-trinorbornane (norbornane) (iii)

(b) Additional Base Groups. Six-Character Abbreviation. The last two digits designate the number of hydrogen atoms in ring. At present, more than 300 base groups are allowed in CHEMOGRAM, provided that their three-dimensional coordinates are given in the database (in STERIC) or their physical properties, observed or estimated, are stored in the database (in EMPRIC). Some examples are as follows.

N06417	7,8-didehydro-4,5-epoxymorphinan (STERIC) N06417(.3.−OH)(.6..A.−OH)(.17.−CH3) morphine morphine is its 3,6 α-dihydroxy-17-methyl derivative
N22208	3*H*-1,4-benzodiazepine (STERIC)

(3) Hetero Groups Symbols. We describe a functional group in nomenclature with a hetero group symbol that is written by either its rational formula or abbreviation. Table 1 shows a classified list of examples, and a full set is shown in the Appendix.

(a) Functional Groups. Hetero group symbols by rational formulas are described only with atom symbols, multiple-bond symbols (at the end(s) and inside), and parentheses.

We selected a variety of functional groups mainly from the "List of Radical Names" at the end of section C and descriptions of organometallic functional groups in section D of the IUPAC Nomenclature, 1979,[13] and gave each of them hetero group symbols. In that list, however, we excluded the following groups: (i) hydrocarbon radicals such as propyl, because they are described by hydrocarbon chains, (ii) hetero atom(s) connected to hydrocarbon chains such as acetyl, because they are described as hydrocarbon and hetero groups, and (iii) radicals including ring(s) such as phenyl, because they are described as derivatives of parent rings.

We classify the symbols into three *direction types*, because types play an important role in chemical logic.

(i) Ordinary Direction Type. This consists of terminal or ending symbols having only a point of attachment to be described at the end of a chain structure as well as two points of attachment to be described in the middle of a chain structure. E.g.,

terminal: OH (−OH) COOH (−COOH)

middle:
O (−O−) OO (−OO−) COO (−COO−) =N (=N−)

(ii) Reverse Direction Type. This consists of a *nonsymmetrical* symbol having two points of attachment, where nonsymmetrical means that atomic arrangement or terminal bonds at the end of a hetero group symbol is not symmetrical. They may appear only in the middle of a string and in reverse of the direction to ordinary direction type. E.g.,

OOC (−OOC−) to COO (−COO−)
N= (−N=) to =N (=N−)

No reverse direction type, however, exists for −OO− because it is symmetrical.

(iii) Initial Direction Type. Symbols are described only at the beginning of a formula. E.g.,

HO (HO−) HOOC (HOOC−) O (O<)

Ordinary and reverse and ordinary and initial direction types are mutually inclusive for a hetero group. But, initial and reverse direction types are mutually exclusive, because we cannot describe, at the beginning of a string, a symbol having points of attachment at both sides, e.g., COO. Therefore, a hetero group can have only a combination of direction types, either ordinary and reverse or ordinary and initial.[20] Designation of either ordinary or reverse direction type for a symbol is rather arbitrary; usually we choose the

Table 1. Examples of Heter Group

direction			formula[a] and functional group name[c]	
ordinary	reverse	initial		
(a) Functional Groups				
(1) Hetero Groups Derived from Water, H_2O				
OH		HO	−OH.	hydroxy (prim., term.)
O		O	−O−	oxy (sec.)
(2) Hetero Groups with One Nitrogen Atom, Derived from Ammonia (Azane), NH_3				
(2-1) Single-Bonded				
NH2		H2N	$-NH_2$	amino (prim., term.)
NH		HN	−NH−	amino (sec.)
N		N	−N<	amino (tert.)
(2-2) Double-Bonded				
=NH		HN=	=NH	imino (sec.)
=N	N=		=N−, −N=	imino (tert., nonsym)
(3) Hetero Groups with Two Nitrogen Atoms				
(3-1) Derived from Hydrazine (Diazane), H_2NNH_2, Single-Bonded				
NHNH2		H2NNH	$-NH-NH_2$	hydrazino (prim., term.)
&NNH2[b]		H2NN	$>N-NH_2$	hydrazino (sec.)
NHNH			−NN−NH−	hydrazo (sec., sym)
NHN	&NNH[b]		−NH−N<, >NNH−	hydrazo (tert., nonsym)
&NN[b]			>N−N<	hydrazo (quat., sym)
(3-2) Derived from Hydrazine (Diazane), H_2NNH_2, Double-Bonded				
=NNH2		H2NN=	$=N-NH_2$	hydrazono (sec., term.)
=NNH	HNN=		=N−NH−, −NH−N=	hydrazono (tert., nonsym)
=NN	&NN=[b]		=N−N<, >N−N=	hydrazono (quat., nonsym)
(3-3) Derived from Diazene, HN=NH				
N=NH		HN=N	−N=NH	azo (prim., term.)
N=N			−N=N−	azo (sec., sym)
(3-4) Miscellaneous				
=NN=			=N−N=	azino (prim., sym)
=N2		N2=	=N=N	diazo (prim., term.)
(4) Hetero Groups with Three Nitrogen Atoms, Derived from Hydrogen Azide (HN3)				
N3		N3	−N=N=N	azido (prim., term.)
(5) Hetero Groups with Nitrogen and Oxygen Atoms				
(5-1) Derived from Hydroxylamine, NH_2OH, Single-Bonded				
NHOH		HONH	−NH−OH	hydroxyamino (prim., term.)
N(OH)		HON	−N(OH)−	hydroxyamino (sec., sym)
ONH2		H2NO	−O−NH2	aminoxy (prim., term.)
NHO	ONH		−NH−O−	aminoxy (sec., nonsym)
&NO[b]	ON		>N−O−	aminoxy (tert., nonsym)
(5-2) Derived from Hydroxylamine, NH_2OH, Double-Bonded				
=NOH		HON=	=N−OH	hydroxyimino (sec., term.)
=NO	NO=		=NO−, −NO=	hydroxyimono (tert., nonsym)
(5-3) Derived from Hydroxylamine *N*-Oxide, $H_2N(\rightarrow O)OH$				
=N(O)OH		HO(O)N=	$=N(\rightarrow O)OH$	aci-nitro (sec., term.)
=N(O)O	ON(O)=		$=N(\rightarrow O)O-$, $-ON(\rightarrow O)=$	aci-nitro (tert., nonsym)
(6) Hetero Groups with Oxygen Atoms, Derived from Formic Acid, HCOOH				
COOH		HOOC	−COOH	carboxy (prim., term.)
COO	OOC		−COO−, −COO−	carboxylato (sec., nonsym)
(7) Hetero Groups with Oxygen and Halogen Atoms, Derived from Formyl Chloride, HCOCl				
COCl		ClOC	−COCl	chloroformyl (prim., term.)
(b) Abbreviation				
(b-1) Conventional Functional Group				
ET		ET	$-C_2H_5$	ethyl (term.)
T.BU		T.BU	$-C(CH_3)_3$	*tert*-butyl (term.)
OAC		ACO	$-OCCH_3$	acetyl (term.)
OCCH3		CH3CO	$-OCCH_3$	acetyl (term.)
(b-2) Amino Acid Residues (Ordinary Direction Only)				
ALA		$-NH-CH(CH_3)-CO-$		L-alanine residue (the middle unit)
H−ALA−		$NH2-CH(CH_3)-CO-$		L-alanine residue (the left unit)
−ALA−OH		$-NH-CH(CH_3)-COOH$		L-alanine residue (the right unit)
H−ALA−OH		$NH2-CH(CH_3)-COOH$		L-alanine

[a] Rational formulas of ordinary and reverse types, if any, are shown. [b] This symbol is unallowable to describe in a CHEMO formula, because usually symbols having more than one point of attachment at the left side (denoted by &) cannot properly be described in line notation. [c] Prim., sec., tert., quat., sym, nonsym, and term. denote primary, secondary, tertiary, quaternary, symmetric, nonsymmetric, and terminal, respectively.

Table 2

b=	direction	no. of points	connectivity no.	symmetry	combination	examples
0	ordinary	1	1		7	−NH2, =NH
1	ordinary	1	≧2	symmetric	8	−NH−
2	ordinary	1	≧2	nonsym	5	=N−
3	ordinary	2	≧2	symmetric		−NHNH−
4	ordinary	2	≧2	nonsym	6	−CONH−
5	reverse	1	≧2	nonsym	2	−N=
6	reverse	2	≧2	nonsym	4	−NHCO−
7	initial	2	1		0	H2N−
8	initial	1	≧2	symmetric	1	HN<

ordinary symbol so that it has carbon or nitrogen atom or a multiple bond at its left side.

Let us explain why we need this discrimination of *direction* in line notation. For example, phenyl acetate, CH_3−COO(ord.)−C_6H_5, has different physical properties from its isomer, methyl benzoate, CH_3−OOC(rev.)−C_6H_5,[20] where the only difference between them is which atom of C and O is connected to the phenyl radical. Only the description following the direction type can discriminate the difference.

Secondly, one can classify hetero groups as primary through quaternary by the number of deleted hydrogen atom(s) from its original compound. We show some examples of hetero group symbols that can exist logically as derivatives of specified compounds in all direction types and classes in Table 1. All symbols of the adopted functional groups have been defined in CHEMO notation.

(b) Abbreviation. We can also describe a variety of functional groups with abbreviations conventional among chemists, all in capital letters; they involve hydrocarbon radicals, too.

(b-1) Conventional Radicals. We adopted only symbols of terminal and initial direction types in line with chemists' practice. E.g.,

PR (propyl), ISOPR (isopropyl), VI (vinyl)

(b-2) Amino Acid Residues. We adopted symbols of 22 L-amino acid residues under direct genetic control (Table 1 in literature[22]) in CHEMO notation. A three-character symbol (all in capitals) designates the middle unit of an acid. Exceptionally, four of them contain ring structures. Only ordinary direction types are allowable. E.g.,

GLN (glutamine)

(4) Inorganics. Inorganics designate inorganic substances and simple substances; exceptionally, an atom count may follow element symbols.

H2	hydrogen molecule
O2	oxygen molecule
H2O	water
CO2	carbon dioxide
CaCO3	calcium carbonate

2. Modifiers. We can designate almost all specifications of constitution and stereochemistry in nomenclature with modifiers. They are written by alphanumerals (all in capital) between two periods. We can also describe numerical values including real ones after some modifiers.

(1) Attributes of Atoms. An attribute of an atom, of which examples are shown below, is usually described just after the designated atom symbol.

(a) Ion: followed by an optional digit; unity may be neglected:

.+. , .+.2	cation
Fe.+.2	iron(II) cation
.−. , .−.3	anion

(b) Isotope:

.I.	isotope (mass number undefined)
.I.*n*	isotope (mass number *n*)
H.I.3	tritium

(2) Attributes of Compounds. An attribute of a compound, of which examples are shown below, is usually described just after the designated molecule or inorganics. They may appear in EMPRIC.

State of aggregation and modifications:

.GAS. or .G.	gaseous phase
.RED.	red modification

(3) Reformation of Parent Rings and Substituents. Usage of them will appear in CONVENTIONS.

(a) Atom replacement;

.AZA.	replacement of a ring atom (any species) by nitrogen

(b) Locant designator in parent ring: Proper usage of numeral types and PS types will appear in CONVENTIONS.

.3.	locant 3.
.3a. or .3A.	locant 3a. (exceptionally, a lower-case is allowed)
.3′. and .3a′. or .3A′.	locants 3′ and 3a′.
.PS.*n*	locant *n*
.PSA.*n*	locant *n*a
.PS.*n*′ or .PSA.*n*′	locant *n*a′ or *n*a′

(4) Geometry and Stereochemistry. Proper place of modifiers will appear in CONVENTIONS.

(a) Coordination around an atom:

.SP.	square planar
.TR.	trigonal pyramid

(b) Configuration around an atom or a bond:

.C. and .T.	cis/trans of a double bond (relative configuration)
.SYN. and .ANT.	syn/anti of a bridged ring (absolute and relative configuration)
.EQ. and .AX.	equatorial/axial (absolute configuration)
.EN. and .EX.	endo/exo (absolute configuration)
.R. and .S.	rectus (right)/sinister (left) (absolute configuration)
.ZUS. and .ENT.	zusammen/entgegen around a double bond (absolute configuration)

(c) Relative configuration around a ring:

.C. and .T.	cis/trans between substituents to a parent ring (relative configuration)
.A. and .B.	alpha/beta (relative configuration)

(d) Conformation around a single bond.

.SP.	syn-periplanar −30° to 30°
.−AC.	anti-(−)clinal −150° to −90°

(e) Conformation of a parent ring.

.CIS.	cis conformer

(5) Other Modifiers. Besides modifiers shown above which are in close connection with the nomenclature, CHEMO notation allows many other modifiers. They are to describe, for example, numerical geometry in STERIC or a value of a physical property in EROICA, or to command a certain processing to the running program. The following shows very few examples; the remainder will appear in Appendix, and the corresponding programs will explain their usage.

.DI.*r*	bond distance of *r* (in Å) (in STERIC)
.VAL.*r*	value *r* of a certain physical property (in EROICA)
.STC.	a command to check stoichiometric equality between reactant and product of a chemical equation (in EROICA)

3. Delimiters. Asterisk (*), equal symbol (=), plus sign (+), and slash (/) serve as delimiters to describe a molecule or an equation. The following examples will explain the function of the delimiters. Two asterisks before and after the compound formula denote a molecule or a mole. A numeral before a formula shows the stoichiometric coefficient including fraction. Numeral 1 for one molecule or mole is compulsory. The last additional asterisk denotes the end of an equation.

1*BENZ(.1.−CH3)(.3.−F)*

(a molecule in an equation)

1*CH2=CH−CH−CH3*+1*NH3*+3/2*O2*=
1*CH2−CH−CN*+3*H2O**

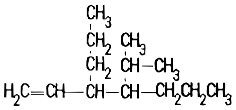
(a reaction equation of acrylonitrile
synthesis by SOHIO process)

CHEMO NOTATION CONVENTIONS

Describing of a complete CHEMO rational formula requires appropriate connection of substance symbols by bond symbols and insertion of requested modifiers in proper positions. As in the conventional rational formula, parentheses denote branches and substituents. The rules in CHEMO notation involve more items than those of SMILES notation because of the higher level as a language of the former. Examples given below show one or a few of valid representations because CHEMO notation never requests *uniqueness*.

1. Fundamentals. (1) Bonds. A bond symbol needs to be described in general *just* before a substance symbol;[23] single bonds are indispensable. Usually, aromatic bonds may not be described because the system detects aromaticity automatically; allyl radical described as −CH2−CH=CH2 may be converted by option to its resonance structure, −CH2,−CH,−CH2.

(2) Branches. Branches specified by enclosures in parentheses can be nested, or stacked,without any limit. The following shows examples to describe 4-isopropyl-3-propyl-1-heptene, where all of them shown below and even some other formulas are valid in CHEMO notation.



CH2=CH−CH(−CH2−CH2−CH3)
−CH(−CH(−CH3)−CH3)−CH2−CH2−CH3

CH2=CH−CH(−PR)−CH(−ISOPR)−PR
PR = propyl, ISOPR = isopropyl

PR−CH(−VI)−CH(−ISOPR)−PR
VI = ethenyl (vinyl)

(3) Repeat Count After Parentheses. A repeat count after a right parenthesis specifying the end of a branch may serve to denote multiple substitution of hydrogens.

N(−C2H5)3 N(−ET)3	triethylamine
O(−C(−CH3)2−CH3)2	T.BU−O−T.BU O(−T.BU)2 di-*tert*-butyl ether T.BU = *tert*-butyl

2. Substituents and Compounds with Two or More Base Groups. Substituents. Chain structures in parentheses after base group symbols represent substituted ring structures; this representation may be the extension of radicofunctional nomenclature such as phenyl chloride. Describing a locant is indispensable, usually just after the right parenthesis; modifiers, if any, may be described just after the locant. As an only exception, .C. (cis) and .T. (trans) may be placed before the locant, as in the case of nomenclature. Repeat count of substituents is allowed.

BENZ(.1.−CH3)(.3.−F)	3-fluorotoluene
DECL(.4a.−CH3)	4a-methyldecahydronaphthalene
CYHA(.1.−ET)(.C..3.−F)	1-ethyl-*cis*-3-fluorocyclohexane
CYHA(.1..AX.−ET)	1(*axial*)-ethylcyclohexane (a less stable structure)
CYHA(.1.−Cl)2	1,1-dichlorocyclohexane

A CHEMO formula can involve *multiple*, that is, two or more, base groups as a matter of course. The second and following base groups are described in a substituent of the preceding base group. A bond before the succeeding base group is placed exceptionally before its locant.

BENZ(.1.−CH2−CH2−.1.BENZ)	1,2-diphenylethane
CH2(−.1.BENZ)−CH2(−.1.BENZ)	1,2-diphenylethane

A *base group domain* denotes a set of a base group and its substituents; an intramolecular linkage may connect the successive base group domains to describe a *multiple base group compound*.

BENZ(.1.−CH2−CH2(−X1)) BENZ(.1.−Z1)	1,2-diphenylethane

3. Short-Cut (Abbreviated) Representation. Some conventional descriptions allow short-cut representation of hydrocarbon chains, besides the hetero group symbols for them.

(1) Normal Alkane. C*n*H*m* (*m* = 2*n* + 2). C16H34 hexadecane.

(2) Normal Alkyl Radical. C*n*H*m* (*m* = 2*n* + 1). C16H33 1-hexadecyl (initial and terminal).

(3) Polymethylene Radical. C*n*H*m* (*m* = 2*n*) or (CH2)*n* C16H32 (CH2)16 1,16-hexadecanediyl(middle).

HO−C16H32−OH or HO−(CH2)16−OH
1,16-hexadecanediol

(4) Polymethine Radical. C*n*H*n* (*n* = even), (CH)*n* (*n* = even) C8H8 (CH)8 1,3,5,7-octatetraene-1,8-diyl (middle).

Cl−C8H8−Cl or Cl−(CH)8−Cl
1,8-dichloro-1,3,5,7-octatetraene

(5) Oligomer. (−Monomer)*n*. *n* = degree of polymerization (middle).

CH3−(−O−CH2)2−Cl equivalent to
CH3−O−CH2−O−CH2−Cl

Oligomers whose monomer contains ring structure(s) can also be described.

4. Ring. A parent ring in a CHEMO formula is described by (1) a CHEMO base group symbol alone, (2) a base group symbol reformed by atom replacement, etc., (3) synthesized base group symbols, or (4) a chain structure ring-closed with intramolecular linkage. Some examples with illustration will be shown also in Results and Discussion.

(1) Reformation by Ring Atom Replacement.

BENZ .AZA. .PS.1 azabenzene (=pyridine)

.PS.1 denotes that carbon of locant 1 is replaced by nitrogen (.AZA.).

(2) Reformation by Hydro/Dehydro Modifiers.

PRDN(.2.−E+)(.3.−E+) 2,3-dihydropyridine

An empty atom E+ denotes hydrogenation at the atom designated by locant.

(3) Reformation by Homologation.

CYHA .HOM.10 .PS.1 cyclohexadecane

.HOM.10 denotes that 10 methylene groups are inserted after locant 1 (.PS.1).

(4) Synthesis by Fusion.

NAPH(.2. .PS.3 .FUS. − .2. .PS.3. PRDN(.4.−Cl))

4-chloronaphtho[*b*:*b*]pyridine = 4-chlorobenzo
[*g*]quinoline

A ring is described by a virtual multiple base group domain formula. A modifier .FUS. designates fusion between two base groups. Locants .2. and .PS.3 designates that atoms of locant 1 and 2 are fusing in NAPH (naphthalene), and locants .2. and .PS.3 are fusing in PRDN.

(5) Synthesis by Spiro.

CYHA(.1. .SPI. − .1.CYHA) spiro[5.5]undecane

A modifier .SPI. designates spiro formation between two base groups. Locants .1. and .1. designate that atoms of locant 1 and 1 in both base groups are fused to make a spiro compound.

5. Stereochemistry. Qualitative representation of stereochemistry such as coordination, configuration, and conformation are described with stereochemical modifiers shown earlier. Both of absolute and relative configuration and conformation can be described as in nomenclature. Quantitative representations such as bond length, bond angle, and dihedral angle as well as coordinates of a specified atom and so on are described with modifiers followed by a real number designating the local or global geometry as shown above. Detail will be shown with examples in a forthcoming paper of STERIC.

6. Other Features. (1) Tautomers. There is no tautomeric bond or normal bond (in CAS INDEX Name[24]) in CHEMO notation. Tautomer structures, e.g., the enol and keto forms, should be discriminated in CHEMO notation because prediction of geometry or properties is only possible with strictly defined structures. In fact, the actual substance is a mixture of tautomers, whose composition depends on the relative stability determined by their free energy of formation at the given condition. Some properties may be of in-between value as mixture, but their geometries are of their own. Therefore, in CHEMO notation, each of the tautomers must be described following their individual structure, whereas a notation such as CAS INDEX Name that primarily aims identification and retrieval of compounds may adopt the specified notation of tautomers. A problem, however, still remains; for better retrieval, the description in the database as well as in query may be duplicated so that all of possible existing structures are retrieved.

(2) Racemic Compound. A racemic compound is certainly a mixture of plural stereoisomers of the same constitution. Therefore, no specified description of racemic compound as a whole exists; each isomer may be described.

(3) Unspecified Configuration. Absence of designation for cis/trans around a double bond shows an example of unspecified configuration. When no discrimination is described of configuration in a given compound, the result we can obtain from the notation depends on programs that receives it. In CHEMOGRAM, EMPRIC gives averaged value of physical properties of cis and trans isomers, whereas STERIC assumes an unspecified isomer as a trans isomer, usually a more stable one.

(4) Molecular Compound. A molecular bond (− −) or an ionic bond („) symbol between composing molecules designates a molecular compound. Examples follows:

BENZ(.1.−OH)(.4.−OH) − − PBZQ	quinhydrone (PBZQ = *p*-benzoquinone)
BENZ(.1.−O.−.) „ Na.+.	sodium phenoxide

STERIC gives coordinates of two molecular components, those of the latter being placed +10 Å apart on the *X*-axis, if no explicit position is described by a modifier. EMPRIC refuses a molecular compound as input, because any sound logic hardly exists for empirical estimation of its physical properties at present.

DECODING OF CHEMO NOTATION

1. Outline. A subsystem DECODE decodes a given CHEMO rational formula into two types of connection tables and some related tables.

(1) Internal Expression of Symbols. The algorithm first recognizes a molecule in a CHEMO input by delimiters and then converts symbols composing a molecule to its *internal expressions*, *ix* and *iy*, both of them being integer. The first internal expression, *ix*, discriminates the major entity or function of a symbol. The second internal expression, *iy*, designates its subsidiary attributes that depend upon the nature of the symbol. Thus, a vector (*ix*,*iy*) shows the internal expression of a symbol. Afterwards, manipulation of symbols in program turns to proceed in numerical processing, apart from literal symbols. Literal symbols appear again in output by reverse conversion of internal expression.

(a) Atom Symbol. *ix* = 1−99. For *ix* = 1 (H) to 92 (U), *iy* = atom count, whereas *iy* = linkage number for intramolecular linkage (*ix* = 96 (X) and 97 (Z)).

(b) Modifier. *ix* = 101−200. *ix* = 101 for bonds and *iy* = 1 for a single bond, 4 for an aromatic bond, etc. *ix* = 111 and 112 for parentheses. *ix* = 120−299 for other modifiers shown earlier. Some examples are .R. (123,21) and .S. (123,22), and .ZUS. (123,91) and .ENT. (123,92); .GAS. (251,1), .LIQ. (251,2), and .SOL. (251,3); .+.2 (256,2), etc. In exceptional cases, *iy* stores a real number, an example of which is .DI.1.24 (131,1.24), that designate bond distance of 1.24 Å.

(c) Hetero Group. *ix* = 301−599. *iy* = *abc* where a. hydrogen deletion (0−4 for primary, secondary, tertiary, and quaternary, respectively), b. direction of the hetero group (0−8. For detail, see NOTE.[25]), c. aldo/keto (1 and 2 for aldo and keto, respectively), which is determined during decoding. Examples are NH2 (307,000), H2N (307,070), NH (307,110) and N (307,210), =NH, aldimine (329,001), and =NH, ketimine (329,002).

(d) Inorganics. *ix* = 601 to unlimited. *iy* is undefined. Examples are H2O (654,0), CO2 (620,0), and CaCO3 (625,0).

(e) Base Group. For a conventional base group, *ix* = −1 to unlimited negative number and *iy* = its number of hydrogen after substitution. Examples are CYHA (−51,-12) and BENZ (−56,6) when unsubstituted. For an additional base group, it will be given as *ix* a temporary negative numeral during decoding.

(2) Conversion of Internal Expressions into Working Tables. After expanding abbreviated symbols always and hetero groups by option, a set of vectors (*ix*,*iy*) represents a molecule, together with vector's ordinals recorded as *internal numbers*, invariant throughout processing. The system analyzes the set from left to right. First, the system divides a molecule into base group domains and records the range of each domain and some other attributes in an auxiliary table. Second, in a base group domain thus divided, the system separates chain structures in an acyclic compound or in substituents and records them in the same way. Third, by detecting the nesting due to various combinations of parentheses in the whole chain structures, a working table stores the set of vectors, connectivity numbers, branching and subbranching atoms, etc. Finally, we convert the working tables of a molecule to its corresponding connection tables.

(3) Connection Tables. Of the two types of connection tables converted from a CHEMO formula, one is a *superatom* (or *atom group*) *connection table*, a network graph of superatoms and atoms as nodes, whose node numbers are cited with *internal numbers*. The atom table consists of superatoms and atoms numbered with discrete original internal numbers,[26] which conserves the order in input. In this table, modifiers are also conserved. The bond table stores bonds between superatoms and atoms. The other is the prevailing *atom connection table* of atoms as nodes, whose node numbers are arbitrarily determined atomic ordinals; base groups as well as hetero groups have been expanded into their composing atom string in an earlier stage of decoding.

The former, the superatom connection table, will facilitate estimating physical properties, the standard enthalpy of formation in gaseous phase as example, by group contribution methods, because it still conserves the essential attributes of superatoms. Both of base groups as well as hetero groups of different classifications due to degree of hydrogen atom deletion (primary, etc.) have their specified contribution values; the different points of attachment of a hetero group (ordinary or reverse) give their characteristic correction contribution.[20] Indeed, only the superatom connection table conserving the concept of superatom as well as the direction of superatoms among them can realize chemical logic concerned with estimation of physical properties.

The latter, the atom connection table, will serve in EROICA and STERIC. In EROICA, a subsystem CANON prepares a canonical representation of a molecule for retrieval. In STERIC, three-dimensional coordinates are calculated from the atom connection table of chain structure by chemical logic, and those of base groups are retrieved from a database. By option, geometry of specific hetero groups can be adopted from a database. A high-level atom connection table is also prepared that contains all the stereochemical descriptions given in a CHEMO formula without any loss of information; this table may be used in some other systems that require stereochemical description.

2. Detection of Aromaticity. In CHEMO notation, base group symbols are associated with a flag whether its specified atom belongs to aromatic ring or not. If an atom string ring-closed by an intramolecular linkage is given, we have to detect aromaticity in the string.

Seemingly, there has been no unique definition of aromaticity, because aromaticity has a variety of implications: equality of contribution in estimating physical properties, symmetry in geometry, reactivity, ring current, etc. Among a few definitions of aromaticity, we have assumed as aromatic a ring or a subring composed of alternative double bonds the number of which satisfies Hückel's ($4N + 2$) rule. Criteria of this pragmatic definition came from the two evaluations, i.e., the above mentioned equality of contribution and symmetry in geometry. The homologues of benzene, i.e., naphthalene, anthracene, phenanthrene, etc. as well as those molecule whose carbon atom(s) are replaced by hetero atom(s), e.g., pyridine, quinoline, isoquinoline, etc., are aromatic, while cyclopentadiene, furan, pyrole, thiophene, tropone, tropylium, etc., are not aromatic as they are not composed of *three* alternative double bonds. Pyridine-*N*-oxide is also aromatic because we can describe it with a Kekulé-like diagram. One of few exceptions is cyclopentadienyl radical, where five carbon atoms are strictly equivalent in all sense of chemistry.

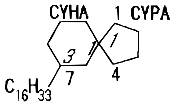
Algorithm in a subsystem FNDRNG detects only six-membered rings having three alternating double bonds, regardless of atom species and designates them as aromatic. During the processing, however, few exceptional cases shown above are also detected.

RESULTS AND DISCUSSION

1. Examples of CHEMO Notation. The systems EROICA/EMPRIC and STERIC of CHEMOGRAM have successfully employed CHEMO notation in describing organic compounds.

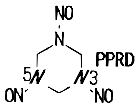
The system EROICA is associated with the EROICA database that stores around 9000 organic compounds. A compound in the database involves three kinds of data terms in a tree structure: (1) Identifiers of a compound: compound name, molecular formula, CAS registry number, CHEMO formula, *individual* formula (see Example 1), and *generic* formula,[27] (2) group contribution count, and (3) values of physical properties.

To demonstrate the performance of CHEMO notation, the following shows a few examples of compounds in the database, only with selected identifiers. At the first phase, the subsystem DECODE decodes a CHEMO formula into an atom connection table.[28] During the processing, base groups had been reformed or synthesized following the designation of the corresponding modifiers. Then a subsystem CANON, whose detail will be presented in a forthcoming paper, prepares the individual formula, a canonical atom-string representation of a compound, from the connection table.



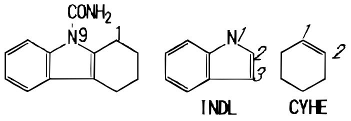
Example 1
name = 7-hexadecylspiro[4.5]decane
CAS REGN = 2307-06-4
CHEMO formula = CYPA(.1. .SPI. −.1. CYHA(.3.−C16H33))
individual formula = CX1X2(C4Z2)C2(C16)C3Z1

A modifier of synthesis, .SPI., and two intrinsic locant descriptors, .1. and .1. (italic in the diagram), represent a parent ring, a spiro compound derived from cyclopentane (CYPA) and cyclohexane (CYHA). Because all carbon atoms in cyclopentane or cyclohexane are equivalent in each ring, any locant number descriptors may be selected in this case. A hexadecyl chain, C16H33 in short-cut description, substitutes hydrogen at the intrinsic locant 3 of the cyclohexane ring. As locant number in a ring is intrinsic in CHEMO notation even after synthesis, the locant 7 (Roman in diagram) in the nomenclature of a spiro compound corresponds to the invariant, intrinsic locant 3 in cyclohexane ring. Then the system derives the atom string from the CHEMO formula, where both rings are expanded in chain structures ring-closed by intramolecular linkages, (X1,Z1) and (X2,Z2), and a branch is inserted in the proper site. Thus, the *individual* formula is a unique and unambiguous string of atoms[29] because it was reasonably canonicalized by a subsystem CANON. Strings of carbon atoms (strictly speaking, in an abridged expression of individual formula) are represented by atom symbols C followed by atom counts, where hydrogen symbols are suppressed. This representation is essentially equivalent to SMILES notation, if one recognizes that X1 or Z1 corresponds to %1, and a long hexadecyl chain would simply be described by C16, instead of the continuous 16 C symbols, which may confuse us in counting.



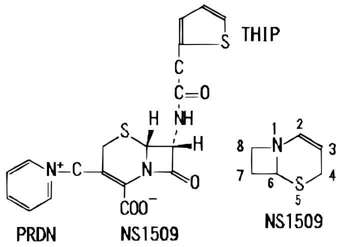
Example 2
name = hexahydro-1,3,5-trinitroso-*s*-triazine
CAS REGN = 13980-04-6
CHEMO notation = PPRD .AZA. .PS.3 .AZA. .PS.5 (.1. −NO)(.3. −NO)(.5.−NO)
individual formula = O=NN(CNX1N=O)CN(CZ1)N=O

Hexahydro-*s*-triazine is a six-membered saturated ring compound with three nitrogen atoms at locants 1, 3, and 5. Therefore, CHEMO notation may describe the parent ring as a ring piperidine (PPRD), two carbon atoms being replaced by nitrogen (.AZA.) at locants 3 and 5 (.PS.3 and .PS.5), respectively;[30] cyclohexane ring replaced by three nitrogen atoms may also be equally allowed. Three inserted nitroso groups (NO) denote substitution at locants shown above. In the individual formula, double bond symbols within nitroso group appear explicitly, while all single bonds are suppressed; it should be noted that an oxygen atom, an atom of highest atomic number, appears at the begin of the individual formula.



Example 3
name = 1,2,3,4-tetrahydro-9*H*-carbazole-9-carboxamide
CHEMO formula = INDL(.2..FUS..PS.3−.2..PS.1 CYHE)-(.1.−CONH2)
individual formula = O=C(N)N(CX1=CX2C4Z1)-QX3QZ2Q4Z3

Fusion of 1*H*-indole (INDL) and cyclohexene (CYHE) introduces a new ring, 1,2,3,4-tetrahydro-9*H*-carbazole. The synthesis modifier (.FUS.) in CHEMO formula represents a fusion of indole at its intrinsic locants 2 (.2.) and 3 (.PS.3), which are connected by a double bond, and cyclohexene at the double bond between its intrinsic locants 2 (.2.) and 1 (.PS.1). As locant 9 of carbazole corresponds to the intrinsic locant 1 of indole, a carboxamide radical (CONH2) substitutes at locant 1 of indole.[31] Symbols "Q" followed by atom counts in the resulting individual formula mean aromatic carbon atoms like "c" in SMILES notation. Careful tracing of the formula will prove the validity of the CHEMO formula.



Example 4
systematic name = 1-{[(6*R*,7*R*)-2-carboxylato-8-oxo-7-{(2-thienylacetyl)amino}-5-thia-1-azabicyclo[4.2.0]oct-2-en-3-yl]methyl}pyridinium
CAS REGN = 50-59-9
trivial name = cephaloridine
CHEMO formula no. 1 (prepared manually) = PRDN-(.1..+.CH2−.3.NS1509(.2.−COO.−.)(.8.=O)(.7..R.−NHCO−CH2−.2.THIP))
CHEMO formula no. 2 (prepared by ENCODE) = NS1509(.2.−COO.−.)(.8.=0)(.3.−CH2(−X1))(.7..R.−NH-CO−CH2(−X2)) THIP(.2.−Z2) PRDN(.1..+.−Z1)
NS1509 = 5-thia-1-azabicyclo[4.2.0]oct-2-ene, cephem
THIP = thiophene

This example from STERIC shows how precisely CHEMO notation can describe a complicated stereospecific compound with aid of a variety of modifiers. The given compound, cephaloridine, has a formidable systematic name in IUPAC nomenclature. CHEMO formula no. 1 shows the formula prepared manually in line with this nomenclature. First, we describe pyridinium, the senior ring system, as PRDN and cation symbols at locant number 1 and methylene in its substituent. Second, we describe NS1509,[32] an additional base group symbol denoting a stereoparent cephem, that is connected at its locant 3 to the methylene. Third, we describe three substituents, i.e., carboxylato anion at locant 2, oxo at locant 8 with a double bond, and a thiophene derivative at locant 7 with a stereomodifier .R. after the locant symbol, in parentheses after their parent ring; (2-thienylacetyl)amino in the name indicates carbamoyl stemming from methylene of 2-methylthiophene, where the amino group in carbamoyl is directly connected to the cephem ring. We need not describe a modifier .R. at locant 6, because the descriptor (6*R*) for cephem has been implied in the stereoparent name and therefore the geometry in the STERIC database of NS1509 has been given to satisfy the requirement.

CHEMO formula no. 2 shows, on the contrary, a formula that the subsystem ENCODE (see below) constructed from its connection table, which has been stored in a certain database. Here, the subsystem constructs the molecule starting at the ring system in which the first atom in the connection table appears. Three base group domains, i.e., those around cephem, thiophene, and pyridine, may be independently prepared in the same way as shown in formula 1. Then they are so connected by two pairs of intramolecular linkages that the aimed compound grows as expected. It may be emphasized that the bycyclic stereoparent as well as stereochemical descripters are automatically encoded by the subsystem.

Geometries calculated by STERIC from two kinds of CHEMO formulas are essentially the same, although coordinates may be shifted or rotated.

2. Usefulness of CHEMO Notation. Here we may summarize the usefulness of CHEMO notation as symbolization of superatoms and as a high-level line notation that is converted from graphical input.

(1) Rings and functional groups have corresponding symbols that are used as universal descriptors in chemical logic-oriented computer programs. (2) Hetero group symbols have the concept of *direction* that is helpful in estimating physical properties.[20] (3) We can convert CHEMO notation into two types of connection tables and *vice versa* by using ENCODE subsystem; a superatom connection table conserves the direction of hetero groups and locants of parent rings. (4) Unambiguous representation of chemical structure including stereochemistry is possible at least in the same level of chemical names following nomenclatures. (5) Rules required in writing line notation is easy for average chemists to study except rare cases. The resulted line notation is readable with ease; the correspondence between the chemical name and the notation may be well judged. (5) The length of notation is much shorter than corresponding connection table. Therefore, it may be used to describe molecules in chemical databases. (6) Unique notation may be prepared by the system CANON.

As CHEMO notation is the higher-order language, however, it invites some unavoidable demerits for chemists to describe it. ENCODE subsystem, prepared by us and to be reported soon, however, may amend the fault, because it can encode a CHEMO formula of a compound without any loss of information from a high-level atom connection table storing even information of stereochemistry. ENCODE discriminates rings, even stereoparents, in the connection table, detects relevant base group symbols in CHEMO notation and furthermore encodes any undetected complicated rings by reformation or synthesis. Absolute and relative stereochemical descriptions in nomenclature including parity representation[12] are also encoded with aid of a subroutine that realizes the sequence rule. It can also encode hetero group symbols from atom strings. CHEMO notation thus prepared indeed can serve chemists as an intermediate language in and as input to the program package CHEMOGRAM.

APPENDIX

The following shows outline of the symbols of atoms, modifiers, hetero groups, inorganics, and base groups. Symbols of additional base groups will be shown when the program STERIC is reported, as they are used mainly there. Lists of full sets of symbols appear in the supporting information. Subsets of symbols may be used if they are selected retaining logical consistency.

Symbols, *ix* and *iy*, and illustrations are given in most lists, whereas in the hetero group list hetero groups classified by *direction* are shown. Statistics and comments follow.

(1) Atom. ordinary atoms, 92; pseudoatoms, 6.

(2) Modifier. bond symbols, 8; parentheses, 2; modifiers, 201.

(3) Hetero Group. No rule approved by chemists seems to be established in writing functional groups in chemical formulas. For instance, carboxy, $-COOH$, is often described as $-CO_2H$; 1,2-diaminoethane, $H_2NCH_2CH_2NH_2$, may be preferred to $\mathit{NH_2}CH_2CH_2NH_2$ concerning the first amino group. Therefore, all possible symbols of a hetero group of a specified type are adopted in CHEMO notation, though seemingly the most prevailing symbols alone are described in the list. Total number of symbols amounts to 339.

(a) Functional Groups. Number of functional groups (a set of groups of all degrees of dehydrogenation is counted as unity.): 80. Total number of symbols: 271 (ordinary: 137; reverse: 28; initial: 106).

(b) Abbreviation. Number of functional groups: 20; total number of symbols: 46 (ordinary: 23; initial: 23).

(c) Amino Acid Residues. Number of functional groups: 22; total number of symbols: 22 (ordinary type only).

(4) Inorganics. Number of inorganics: 193.

(5) Base Group. Number of ordinary base groups: 244 (number of parent hydrocarbon rings: 35; number of parent heterocyclic rings: 86; number of other rings: 123). Structural diagrams of base groups will appear in a paper of STERIC.

ACKNOWLEDGMENT

It is my greatest pleasure to submit a paper to the special issue of this Journal dedicated to celebrate the brilliant achievement of Prof. S. Sasaki; for this quarter of a century I have been his frail comrade. Research on CHEMOGRAM and CHEMO notation was initiated by the author at the Department of Synthetic Chemistry, Faculty of Engineering, the University of Tokyo and completed at the Institute of Research and Development, Tokai University. The author thanks many colleagues for their kind discussion and research aides who help me in practical and clerical works for their faithful assistance.

Supporting Information Available: Lists of the full set of symbols (20 pages). This material is contained in many libraries on microfiche, immediately follows this article in the microfilm version of the journal, can be ordered from the ACS, and can be downloaded from the Internet; see any current masthead page for ordering information and Internet access instructions.

REFERENCES AND NOTES

(1) Lynch, M. F.; Harrison, J. M.; Town, W. G. *Computer Handling of Chemical Structure Information*; MacDonaldo: London, 1971.

(2) *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; Wiley: New York, 1974.

(3) *Chemical Information Systems*; Ash, J. E., Hyde, E. Eds.; John Wiley: New York, 1975.

(4) Cited in ref 3. Dyson, G. M. A Notation for Organic Compounds. *Nature* **1944**, *154*, 114.

(5) Wiswesser, W. J. *A Line-Formula Chemical Notation*; Crowell: New York, 1954; Rev. ver.: MacGraw-Hill: New York, 1968.

(6) Yoneda, Y. *KEMOGURAMU(CHEMOGRAM)*—*Keisanki Kogyokagaku*; Maruzen: Tokyo, 1972; Vol. 1 (in Japanese).

(7) Yoneda, Y. A Proposal of an Estimation and Retrieval System EROICA for Physical Properties of Organic Compounds by CHEMO Inputs. *In Information Chemistry*. *Computer Assisted Chemical Research Design*; Fujiwara, S.; Mark, H. S., Jr., Eds.; University of Tokyo Press: Tokyo, 1975; p 239.

(8) 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.

(9) Araki, K.; Kaji, M. A Stereochemically Accurate Chemical Substance Database Based on the Systematic Names of Organic Compounds. 1. Low Molecular Weight Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 363−374.

(10) McDaniel, J. R.; Balmuth, J. R. Kekulé: OCR—Optical Chemical (Structure) Recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373−378.

(11) Yoneda, Y. Recognition of Chemical Diagram Described with Characters, Preprint, 13th Symposium on Information Chemistry, Division of Chemical Information, The Chemical Society of Japan, Toyohashi, 1990; pp 33−36, 28I08 (in Japanese).

(12) Wipke, W. D.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.

(13) IUPAC Nomenclature of Organic Chemistry, Sections A−F, and H; Pergamon: Oxford, England, 1979.

(14) Yoneda, Y. Empirical Calculation of Three-dimensional Structures of Organic Compounds—STERIC/ENCODE, Preprint, 12th Symposium on Information Chemistry, Division of Chemical Information, The Chemical Society of Japan, Osaka, 1989; pp 50−53, 8I12 (in Japanese).

(15) Yoneda, Y. An Estimation of the Thermodynamic Properties of Organic Compounds in the Ideal Gas State. I. Acyclic Compounds and Cyclic Compounds with a Ring of Cyclopentane, Cyclohexane, Benzene, or Naphthalene. *Bull. Chem. Soc. Jpn.* **1979**, *52*, 1297−1314.

(16) Yoneda, Y. EROICA Database: A Database for Fundamental Physical Properties of Organic and Organometallic Compounds, Proceedings of the 7th International CODATA Conference, Kyoto; Pergamon: London, 1981; pp 254−257.

(17) Yoneda, Y. A Computer Program Package for the Analysis, Creation and Estimation of Generalized Reactions—GRACE. I. Generation of Elementary Reaction Network in Radical Reactions—A/GRACE-(I). *Bull. Chem. Soc. Jpn.* **1979**, *52*, 8−14.

(18) Element symbols described by two capital letters plus an apostrophe (’) are also allowed for two-character atomic symbols. Examples: CL’ and CA’.

(19) Bond symbols belong to modifiers in the grammar of CHEMO notation.

(20) Enthalpies of formation of gas at standard state, $\Delta H^\circ_f(g)$, of isomers: ref 21. $CH_3COOC_6H_5$: phenyl acetate, −279.7 ± 2.4 kJ/mol; $C_6H_5COOCH_3$: methyl benzoate, −287.9 ± 1.2 kJ/mol; CH_3-$COOC_4H_9$: butyl acetate, −485.6 ± 0.7 kJ/mol; $C_4H_9COOCH_3$: methyl pentanoate, −471.2 ± 0.9 kJ/mol. In the latter pair, EMPRIC estimates their $\Delta H^\circ_f(g)$ as butyl acetate [Δ(hexane) + Δ(COO) + Δ(type correction in direction of ∼OOC, due to the secondary carbon atom in C4H8)] and methyl pentanoate [Δ(hexane) + Δ(COO) + Δ(type correction in direction of ∼COO, due to the secondary carbon atom in C4H8)], where Δ indicates contribution values due to symbols or correction in parentheses. The difference between two isomers arise from that of type correction due to *direction* of the COO symbol.

(21) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*, 2nd ed.; Chapman and Hall: London and New York, 1986; pp 130, 133.

(22) IUPAC CNOC; IUPAC-IUB CBN, Nomenclature of Alpha-Amino Acids. Recommendation, 1974. *Eur. J. Biochem.* **1973**, *33*, 1−14.

(23) Exceptionally, a bond symbol to a following base group symbol (BENZ in example) should be described before the locant modifiers (.3.). e.g., CYHA(.1.−.3.BENZ).

(24) Chemical Abstracts Service, *Naming and Indexing of Chemical Substances for CHEMICAL ABSTRACTS*, 1987 Index Guide; American Chemical Society: 1987; 122, 184.

(25) Hetero group symbols of the same degree of dehydration (*a*) will be classified (*b*) according to their direction, number of points of contact, connectivity number and symmetry. Two *b*’s are allowed to a hetero group of the same *a* except for *b* = 3. See Table 2.

(26) Two types of superatom connection tables are prepared by optition. One stores intramolecular linkage symbols (X,Z) as pseudoatoms, and the other does not.

(27) A generic formula means a canonicalized formula like an individual formula but conserves superatoms in preparing the unique string of symbols. *Resemblance isomers*, a set of isomers having resembling physical properties, are represented by a generic formula. Detail will be given in a forthcoming paper on GENERC.

(28) Superatom connection tables are not employed in EROICA database. They will, however, play an important role in estimation in EMPRIC.

(29) At present stage, an individual formula cannot discriminate against stereoisomers such as cis/trans.

(30) Use of many period marks to denote modifiers and locants is certainly cumbersome. The author thinks this inefficiency may be a reluctant compromise between human and machine.

(31) In the CHEMO formula of example 3, a confusion may arise whether the last substituent (.1.−CONH2) belongs to INDL or CYHE. This substituent is of the same level as that of the preceding substituent, (.2..FUS....), therefore belongs to the base group INDL(indole). If the former substituent were connected to CYHE, it might be described just after CYHE, as INDL(.2........CYHE(.1.−CONH2)).

(32) How can anybody find an additional base group symbol NS1509 for cephem? When a chemist is manually describing formula no. 1, he or she may find it in a classified book of printed structural diagrams of base groups. As for formula no. 2, this symbol is automatically selected when a connection table is encoded to CHEMO notation by ENCODE subsystem.

CI9501094