XyM Notation for Electronic Communication of Organic Chemical Structures

Shinsaku Fujita* and Nobuya Tanaka

Department of Chemistry and Materials Technology, Kyoto Institute of Technology,
Matsugasaki, Sakyoku, Kyoto 606-8585, Japan

The XyM notation system is proposed as a linear notation for the electronic communication of structural formulas. Each XyM notation consists of a skeleton with such arguments as SUBSLIST, ATOMLIST, and BONDLIST. The arguments are designed to be capable of carrying out substitution derivation for placing large substituents, atom derivation for spiro fusion, and bond derivation for ring fusion. Additional arguments, SKBONDLIST and OMITLIST, are discussed for stereochemical and ring-opening information. The XyM notation system is implemented as a LaTeX2e application.

1. INTRODUCTION

How should we represent chemical structural formulas in computers? This has continuously been a serious problem for information chemists who have been engaged in the construction and maintenance of chemical databases[1−3] as well as for organic chemists who have aimed at chemical applications of computers.[4,5]

The first type of such representation systems has been linear notations. They were at first concerned with data storage and retrieval by computer, as summarized in excellent reviews[6−9] and books.[10−13] For these purposes, the representation of chemical structures has been required to have a unique and unambiguous nature.[10] In particular, the Dyson−IUPAC notation,[14,15] the Wiswesser line-formula notation (WLN),[16] the Hayward notation,[17] and the Skolnik notation[18] are representatives of such data-handling notations. Later, more flexible approaches have been possible in light of the advances in computer technology. For example, SMILES (simplified molecular input line entry system),[19−21] the (extended) Wilcox−Levinson system,[22,23] and SYBYL line notation[24] have been reported as chemical notation systems, putting more emphasis on simple encoding rules and on clarity of the encoded expressions.

Methods belonging to the second type have been called topological methods, which have mainly been based on the concept of connection tables[25−27] or equivalents.[28] They have been adopted by most database systems and application systems for manipulating chemical structures.[29,30] Canonical numbering and naming by Morgan's algorithm[31] or others[27,32−42] have been reported for creating unique and unambiguous representations on the basis of the topological methods.

The third way of representing the data of chemical structures has been to store the corresponding image data in a binary format. Such binary data or equivalent text data (e.g., Encapsulated PostScript data) have been widely used in desktop publishing techniques (DTP). A review[43] appearing at this stage of progress has concluded that image data of chemical diagrams produced by an interactive structure editor[44] have advantages over command languages[45] which can be regarded as linear notations for chemical structures.

It should be noted, however, that this conclusion has been derived from a viewpoint of DTP application, not from a viewpoint of electronic communication.

The last 5 years of this decade have seen deep influences provided by the progress of electronic communication. Data produced by a DTP system have been transmitted in a digital format from an author to a publisher. Moreover, the worldwide web (WWW) system for the Internet has offered such data for public use by means of browser techniques, where HTML (hypertext markup language) has played an important role.[46] The present progress means that common data in a digital format can be shared by individual systems having various targets (e.g., DTP, electronic publishing, WWW browsing, and database systems). Now, the problem setting described at the top of this section has been changed into another challengeable one: How do we communicate chemical structural formulas between such systems? In other words, a common language for representing structural formulas should be developed in order to realize effective electronic communication. It should have the following characteristics, different from those of languages for storage and retrieval and for traditional (spoken and literary) communication:

(1) Unique nature is unnecessary for communication languages, while unambiguous nature should be maintained. Nonunique nature is, rather, a requisite, since various standpoints may exist, depending upon the aspects of discussion on a chemical structure. In other words, the language permits synonyms for representing a single structure.

(2) The language should be written as an ASCII text or equivalent, which can be easily generated by computers as well as by humans.

(3) The language should be readable and understandable by computers as well as by humans so that it is capable of reproducing a structural diagram easily.

(4) The language should contain a procedure of linking structural units, so as to represent substitution, spiro ring replacement, and ring fusion. The procedure should not depend on layout data.

(5) The language should contain stereochemical information not in the form of layout data (such as x,y,z coordinates)

but in the form of logical designation (such as α and β in the IUPAC nomenclature). The information should be reproducible by computers and by humans so as to depict the corresponding structural diagram easily.

The IUPAC nomenclature,[47,48] the CAS nomenclature,[49] and the HIRN nomenclature[50,51] are represetative traditional communication languages satisfying the first and second requisites, as mentioned previously.[6] However, it is necessary to develop appropriate software for interpreting such languages in order to satisfy the other items.[52−54] We have reported the XyMTeX system for typsetting chemical structural formulas,[55−58] where the XyMTeX commands are regarded as linear notation.[59] Although the XyMTeX command system has been adopted in a WWW browsing system,[60] it requires layout data in complicated cases, which should be avoided so as to satisfy items 4 and 5 described above. Another type of approach in this discipline has been based on the chemical markup language (CML),[61] which has been defined as a superset of HTML. The American Chemical Society has proposed chemical exchange format (CXF),[62] which is based on ASN.1.[63] DTP-oriented languages such as the XyMTeX command system, WWW-oriented languages such as CML, and database-oriented languages such as CXF may be candidates for common languages of representing chemical structural formulas, if they are extended so that they satisfy the items listed above in order to cover the full domain of electronic communication.

As clarified in the preceding paragraphs, our target is the systematization and extension of the XyMTeX command system so that it will serve as a common language for electronic communication, which is called the XyM notation system[64] in this paper. Another target is the implementation of the XyM notation system to cover DTP application, which is one of the important disciplines of electronic communication. Its application to the WWW technique will be discussed in our forthcoming paper.

2. SKELETONS AND THEIR DERIVATIVES

2.1. Derivation Processes. The XyM notation system proposed here is based on the methodology that each organic compound is regarded as a derivative of a mother skeleton, where such derivation processes are categorized into (1) substitution derivation, (2) atom derivation, or (3) bond derivation. This methodology is an extension of the XyMTeX command system reported previously.[59] The three derivation processes are considered to produce a family of derivatives coming from a skeleton, which are collectively called a "derivative family" for the sake of simplicity. A derivative derived from the skeleton \GenSkel is designated by XyM notation of the format \GenSkel[BONDLIST]{ATOMLIST}-{SUBSLIST}, where \GenSkel represents the name of a generic skeleton and each list in braces or brackets is an argument to specify the bonds, atoms, or substituents. These lists are designated in a systematic manner to produce a derivative family from \GenSkel. The word "generic skeleton" is used to call a skeleton that is specified by the three lists (BONDLIST, ATOMLIST, and SUBSLIST). Table 1 shows a sample family of derivatives starting from a six-membered skeleton as well as the corresponding XyM notations, where \sixheterov is selected as a representative of \GenSkel. It should be emphasized here that the XyM

Table 1. Derivation Processes for XyM Notation

derivation process	derivative family produced from \sixheterov	
substitution	\sixheterov{}{2==F}	\sixheterov{}{2Sa==F; 2Sb==F}
atom	\sixheterov{1==O}{}	\sixheterov{1==O; 4==O}{}
bond	\sixheterov[b]{}{}	\sixheterov[ace]{}{}

notation system is regarded as a linear notation for electronic communication after substantial extension, though each XyM notation apparently takes the same format as a XyMTeX command.

The **substitution derivation** is carried out by writing the argument SUBSLIST of the skeleton \GenSkel. For example, the SUBSLISTs in the codes of this derivation (Table 1) contain {2==F} for the XyM notation of fluorocyclohexane and {2Sa==F;2Sb==F} for the XyM notation of 1,1-difluorocyclohexane, where a semicolon is used to delimitate every substitution mode and a double equality symbol is used as a delimiter between a substitution position and a substituent. The locant number of each substitution is fixed for every skeleton, since substitution sites that are regarded as being equivalent in terms of the IUPAC nomenclature should be discriminated for the purpose of electronic communication. Thus, for example, the XyM notation \sixheterov{}-{n==F} with $n = 1−6$ (due to clockwise numbering from the top of the vertical hexagon) generates six distinct but chemically equivalent diagrams of fluorocyclohexane. Each of them has an equal opportunity to be selected as an illustrated diagram.

Each substitution position may be accompanied by a bond modifier, which represents a bond type (e.g., "D" for an exocyclic double bond), the direction of the bond attachment (e.g., "Sa" or "Sb" for a single bond with no specific stereochemisty), or sometimes the stereochemistry of the bond (e.g., "SA" or "SB" for an α or β single bond). The list of such bond modifiers is shown in Table 2. The bond modifiers Sd (d for down) and Su (u for up) designate α and β bonds in such an exchanged manner as the bond modifiers SA and SB designate. Ring fusion requires bond modifiers for designating substitution at such fused positions, i.e., FA, FB, GA, and GB.

Since the directions of exocyclic bonds depend on the substitution positions, they are described by integers 0 to 8 in both the right- and left-hand sides (Figure 1). Then, an appropriate set of bond modifiers with fixed directions is assigned to each position after categorized as v1 to v6 or h1 to h6 (Table 3). The symbols "v" and "h" are associated with the suffixes v and h in the XyM notation, which will be explained in the next section. Therefore, auxiliary information on a bond (multiplicity, stereochemistry, etc.) can be treated in a simple manner. For example, the sets designated by the set identifiers v1 to v6 (Table 3) are assigned to the positions 1 to 6 of a vertical six-membered ring. Figure 2 (left) exemplifies the assignment concerning

XyM Notation for Electronic Communication

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **905**

Table 2. Bond Modifiers for SUBSLIST

bond modifiers	printed or displayed structures
none or S	exocyclic single bond
D	exocyclic double bond
A	α single bond
B	β single bond
Sa	α (not specified) single bond
Sb	β (not specified) single bond
SA	α single bond (dotted line)
SB	β single bond (boldface)
Sd	α single bond (dotted line) with an alternative direction to SA
Su	β single bond (boldface) with an alternative direction to SB
F	single bond without stereochemistry for ring fusion
FA	α single bond (dotted line) for ring fusion
FB	β single bond (boldface) for ring fusion
G	single bond without stereochemistry for ring fusion
GA	α single bond (dotted line) for the other ring fusion
GB	β single bond (boldface) for the other ring fusion
W	terminal bond for zigzag polymethylenes
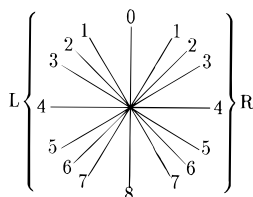
Figure 1. Bond directions for substituents.

v1 to v6 by several illustrations. Thus, the absence of a bond modifier (or S) at position 1 corresponds to a single bond at direction 0, where it is selected from the set v1; bond modifiers Sa and Sb at position 2 generate single bonds at directions R1 and R4 (v2); bond modifier D at position 3 gives a double bond at direction R5 (v3); bond modifiers SA and SB at position 4 provide α and β bonds at directions R7 and L7 (v4); bond modifier A at position 5 yields an α bond at direction L5 (v5); and bond modifier B at position 6 gives a β bond at direction L3 (v6). On the other hand, the six positions 1−6 of a horizontal six-membered ring (produced by \sixheteroh) take the sets designated by the set identifiers h1 to h6 as exemplified in Figure 2 (right).

Atom derivation is the replacement of vertexes by appropriate heteroatoms, where the argument ATOMLIST of the skeleton is used to designate such heteroatoms. The ATOMLIST contains a semicolon for delimitating every replacement mode, each of which in turn contains a double equality symbol for separating a position number from the heteroatom to be replaced. For example, the code {1==O} in the XyM notation of oxane (tetrahydrofuran) indicates that position 1 is replaced by an oxygen atom, and the code {1==O;4==O} in the XyM notation of 1,4-dioxane shows that positions 1 and 4 are replaced by oxygen atoms (Table 1). It should be noted here that the atom derivation proposed for the XyM notation is closely related to the Hantzsch−Widman nomenclature adopted as IUPAC Rule B-1[65] or the replacement nomenclature of IUPAC Rule B-4.[47] Thus, the code {1==O} corresponds to the IUPAC names oxane (Rule B-1) and oxacyclohexane (Rule B-4), while the code {1==O;4==O} is related to the IUPAC names 1,4-dioxane (Rule B-1) and 1,4-dioxacyclohexane (Rule B-4).

Bond derivation is the introduction of unsaturation to a skeleton. The argument BONDLIST of a skeleton contains lowercase alphabets to designate such unsaturation, where

an alphabet is assigned to each bond of the skeleton. Thus, the letter a is given to a bond between positions 1 and 2, b to a bond between positions 2 and 3, and so on. Such designation of bonds comes from IUPAC Rule A-21.5 for naming fused carbocyclic compounds or Rule B-3 for naming fused heterocyclic compounds.[47] As illustrated in Table 1, the code [b] for the XyM notation of cyclohexene indicates that bond b (a bond between positions 2 and 3) is a double bond and the code [ace] for the XyM notation of benzene (a Kekulé structure) shows that bonds a, c, and e have been unsaturated. The BONDLIST argument is optional and may be omitted for fully saturated cases.

The three derivation processes may be freely combined, as shown in Table 4. The combination of substitution and atom derivation is exemplified by 2-fluorothiane and 3,3-difluoro-1,4-oxathiane. 1-Fluorocyclohexene and 3,3-difluorocyclohexene are examples of the combination of substitution and bond derivation. The inspection of the corresponding XyM notations shows the closeness between the XyM notation system and the IUPAC nomenclature. On the other hand, the remaining examples listed in Table 4 show the difference between them. Thus, 3,4-dihydro-2*H*-oxine (dihydropyrane) and pyridine are a case of atom and bond substitution. 5-Fluoro-3,4-dihydro-2*H*-dihydro-2*H*-oxine and 2-chloropyridine are examples of three derivation processes combined. The difference comes from the fact that the XyM notation system adopts a fully saturated molecule as a mother skeleton, while the IUPAC nomenclature selects a most highly unsaturated molecule.

2.2. Generic and Specific Skeletons. Table 5 lists generic skeletons for the XyM notation, each of which at least takes an ATOMLIST, a SUBSLIST, and a BONDLIST (optional argument) as arguments. Table 5 also collects skeletons for giving zigzag polymethylenes. The name of each skeleton consists of a backslash symbol that represents the top of a XyM notation, a stem that comes from the ring size or fusion mode, and a suffix that represents the direction of drawing the skeleton structure (v, vi, vb, vt, h, and hi). The suffix v for a six-membered ring means that the XyM notation prints or displays a structural formula of a vertical hexagon. The suffix h means that the XyM notation generates a structural formula of a horizontal one. The suffix v or vi for a five-membered ring indicates that the five-membered ring is generated by truncating the top or the bottom vertex of a vertical six-membered ring to give a horizontal edge (e.g., \fiveheterovi from \sixheterov and \fiveheterov from \sixheterovi), where the locant numbering starts from the nontruncated bottom or top vertex. When alternative orientations are possible, XyM notations are differentiated by an additional suffix i, b, or t, as found in Figure 3. For example, the suffix vi indicates that the locant numbering starts from an inverse top in the same ring of a skeleton with v and advances in an *inverse* manner (counterclockwise if the v is concerned with a clockwise manner). The suffix vb indicates that locant 1 is assigned to an upper vertex of the *bottom* six-membered ring of a fused ring as found in the third formula of Figure 3, while the suffix vt indicates that the locant numbering starts from an upper vertex of the *top* six-membered ring, as shown in the fourth formula of Figure 3.

Straight-chain polymethylenes can be drawn by using the di- to decamethylene skeletons listed in the bottom of Table 5. A pair of skeleton names with and without the suffix i

Table 3. Sets of Bond Modifiers

set ID	R or L	bond direction[a]						
		0	1	3	4	5	7	8
v1	R	S, D, A, B	Sa, SA, Su	G, GA, GB				
	L		Sb, SB, Sd	F, FA, FB				
v2	R	F, FA, FB	Sb, SB, Sd	S, D, A, B	Sa, SA, Su	G, GA, GB		
v3	R		F, FA, FB		Sa, SA, Su	S, D, A, B	Sb, SB, Sd	G, GA, GB
v4	R				F, FA, FB	Sa, SA, Su	S, D, A, B	
	L				G, GA, GB	Sb, SB, Sd		
v5	L			G, GA, GB	Sa, SA, Su	S, D, A, B	Sb, SB, Sd	F, FA, FB
v6	L	G, GA, GB	Sb, SB, Sd	S, D, A, B	Sa, SA, Su	F, FA, FB		
h1	L		G, GA, GB	Sb, SB, Sd	S, D, A, B	Sa, SA, Su	F, FA, FB	
h2	R	Sa, SA, Su	G, GA, GB					
	L		S, D, A, B	Sb, SB, Sd	F, FA, FB			
h3	R	Sa, SsA, Su	S, D, A, B	Sb, SB, Sd	G, GA, GB			
	L		F, FA, FB					
h4	R		F, FA, FB	Sb, SB, Sd	S, D, A, B	Sa, SA, Su	G, GA, GB	
h5	R				F, FA, FB	Sb, SB, Sd	S, D, A, B	Sa, SA, Su
	L					G, GA, GB		
h6	R						F, FA, FB	SA, SA, Su
	L				G, GA, GB	Sb, SB, Sd	S, D, A, B	

[a] See Figure 1. Directions 2 and 6 are not used here.

Table 4. Combined Derivation Processes for XyM Notation

derivation process	derivative family produced from \sixheterov	
substitution and atom	\sixheterov{1==S}{2==F}	\sixheterov{1==S; 4==O}{2SA==F; 2SB==F}
substitution and bond	\sixheterov[b]{}{2==F}	\sixheterov[c]{}{2Sa==F; 2Sb==F}
atom and bond	\sixheterov[b]{1==O}{}	\sixheterov[ace]{1==N}{}
substitution, atom, and bond	\sixheterov[b]{1==O}{3==F}	\sixheterov[ace]{1==N}{2==Cl}
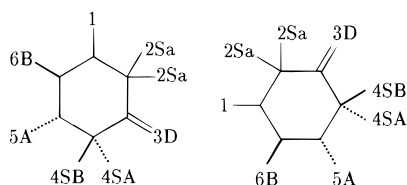
Figure 2. Examples of locant numbers and bond modifiers.

designates the downward or upward stroke of a zigzag chain at the left-hand terminal (see Figure 10 explained later). Branched-chain polymethylenes are depicted by the combination of the di- to decamethylene skeletons by virtue of substitution or atom derivations. Since the XyM notation system does not aim at unique nature, its rules for alkane nomenclature are not always parallel with the IUPAC rules based on the longest chain and alphabetical ordering substituents.
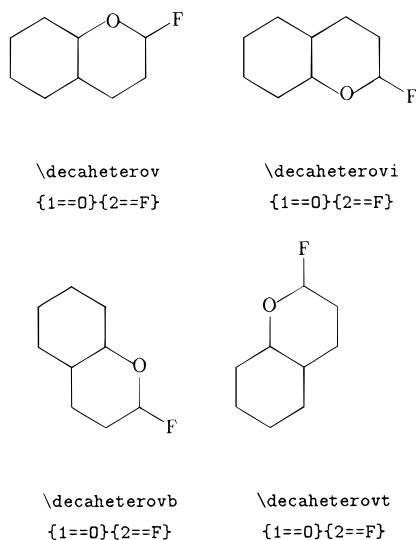
On the other hand, a specific skeleton \SpecSkel is defined as a skeleton with a fixed ATOMLIST: \SpecSkel-[BONDLIST]{SUBSLIST}, where the BONDLIST and SUBSLIST have the same meaning as those of generic skeletons \GenSkel. Since an ATOMLIST is given im-

plicitly, the resulting derivatives are involved in a family of derivatives of the corresponding \GenSkel as a sub-family. Table 6 lists carbocyclic skeletons, each of which is generated by placing carbon atoms at every vertex of the corresponding skeleton listed in Table 5. For example, XyM notation based on the specific skeleton, \cyclohexanev-[BONDLIST]{SUBSLIST}, has the same meaning as the one generated from the generic skeleton, \sixheterov-[BONDLIST]{}{SUBSLIST}, where the latter ATOMLIST is vacant to indicate an implicit set of carbons (Figure 4). Table 7 lists specific skeletons for giving N-containing heterocycles. Figure 4 also shows the comparison between \pyridinev and \sixheterov, where the former lacks an ATOMLIST because of the implicit designation of a nitrogen atom. Other specific skeletons can be defined in a similar way on the basis of the generic skeletons listed in Table 5.

Since carbocyclic aromatic compounds appear frequently in the chemical literature, their corresponding specific skeletons are defined as \CaromSkel[BONDPATTERN]-{SUBSLIST}, where BONDPATTERN represents a bond pattern for aromatic or related unsaturation (e.g., r for a right-hand Kekulé formula, l for a left-hand Kekulé formula, and

Table 5. Generic Skeletons (\GenSkel)

ring size	generic skeleton for cyclic compounds			
3	\threeheterov	\threeheterovi	\threeheteroh	\threeheterohi
4	\fourhetero			
5	\fiveheterov	\fiveheterovi	\fiveheteroh	\fiveheterohi
6	\sixheterov	\sixheterovi	\sixheteroh	\sixheterohi
5,6	\nonaheterov	\nonaheterovi	\nonaheteroh	\nonaheterohi
6,6	\decaheterov	\decaheterovi	\decaheteroh	\decaheterohi
	\decaheterovb	\decaheterovt		

length	generic skeleton for polymethylenes			
2 and 3	\dimethylene	\dimethylenei	\trimethylene	\trimethylenei
4 and 5	\tetramethylene	\tetramethylenei	\pentamethylene	\pentamethylenei
6 and 7	\hexamethylene	\hexamethylenei	\heptamethylene	\heptamethylenei
8 and 9	\octamethylene	\octamethylenei	\nonamethylene	\nonamethylenei
10	\decamethylene	\decamethylenei		

\decaheterov
{1==O}{2==F}

\decaheterovi
{1==O}{2==F}

\decaheterovb
{1==O}{2==F}

\decaheterovt
{1==O}{2==F}

Figure 3. Suffixes for vertical 6-to-6 fused rings.

c for an aromatic circle). Table 8 lists specific skeletons for giving such aromatic compounds. The BONDPATTERN argument is optional; hence, the default pattern is generated when the argument is omitted. Figure 5 illustrates synonymous XyM notations for 1-fluoronaphthalene by using \decaheterov (generic skeleton), \decalinev (specific one), and \naphdrv (carbocyclic aromatic one). A comparison among them shows that the designation of arguments is simplified in the order of \GenSkel, \SpecSkel, and \CaromSkel.

For complicated skeletons with conformation or other additional features, more skeleton names should be added. Their implementation is one of the targets of future investigations. The present paper, rather, is devoted to the extension of a variation of derivation patterns, as found in the next sections, by starting from the skeletons listed in Tables 5−8.

3. EXTENSION OF DERIVATION PROCESSES

3.1. Basic Ideas for Extension. To represent a wider range of chemical structures, the three derivation processes are extended so as to treat complicated substitution, ring fusion, and spiro replacement. Therefore, a derivative family is expanded to accommodate such derivatives created by the extensions. The basic idea is summarized in Table 9, which shows additional derivatives of the derivative family generated from \sixheterov (see Tables 1 and 4).

3.2. Substitution Derivation for Large Substituents. Any derivative derived from a skeleton (\GenSkel or \SpecSkel) can be converted into the corresponding substituent by adding the code (yl) with a locant number (*n*) to its SUBSLIST: \GenSkel[BONDLIST]{ATOMLIST}{...;*n* ==(yl);...}; \SpecSkel[BONDLIST]{ATOMLIST}{...;*n* ==(yl);...}. The resulting notation for a substituent can be added to the SUBSLIST of any other skeleton so as to give a final notation that contains the full connectivity data of the combined structure. For example, the skeleton \sixheterov is converted into a substituent (cyclohexyl) by writing \sixheterov{}-{5==(yl)}, as found in the first example of Table 9. Then, this is added to the SUBSLIST of the outer skeleton \sixheterov, giving cyclohexylcyclohexane. This technique for generating a substituent is called "yl function".

Two or more substituents generated by the yl function can be introduced into a SUBSLIST. The structural formulas of terphenyl and hexaphenylbenzene can be drawn by this technique, as shown in Figure 6. Note that a % symbol indicates the terminal of each line to avoid an extra space in the XyM notation. Although multiplicative terms may be helpful in communication, they are not supported by the present XyM notation system, as found in the code for hexaphenylbenzene. This is because our effort has been devoted to the simplicity of computer processing.

Two or more yl functions can be nested. For example, 4-(4-benzoylcyclohex-3-en-1-yl)-1-naphthol is drawn by the XyM notation

\decalineh[acfhk]{1==HO;4==%

\cyclohexaneh[c]{1==(yl);4==%

\tetrahedral{2==(yl);0==C;1D==O;
$\qquad$ 4==\cyclohexaneh[ace]{1==(yl)}}}}

where we carry out such a nesting as (outer) \decalineh ← \cyclohexaneh ← \tetrahedral ← \cyclohexaneh (inner). Note that \tetrahedral is capable of representing a carbonyl group by adding the code 1D==O in its subsitution list.

The yl function provides us with a versatile tool to generate a substituent that is linked directly to a substitution site of a mother skeleton. There are, however, many cases in which a substituent is linked to a substitution site through an intervening unit (e.g., O, SO₂, and NH). The intervening skeleton \ryl is used to generate a right-hand substituent with

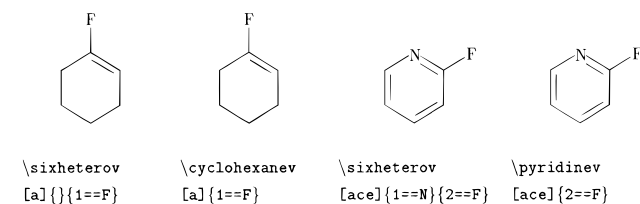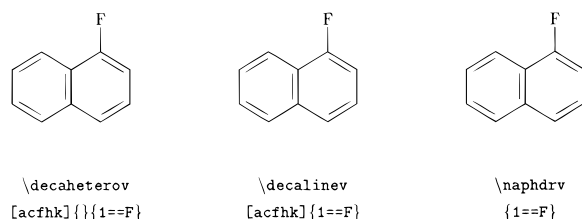**Table 6.** Specific Skeletons for Carbocyclic Compounds (\SpecSkel)

ring size	carbocyclic skeleton			
3	\cyclopropanev	\cyclopropanevi	\cyclopropaneh	\cyclopropanehi
4	\cyclobutane			
5	\cyclopentanev	\cyclopentanevi	\cyclopentaneh	\cyclopentanehi
6	\cyclohexanev		\cyclohexaneh	
5,6	\indanev	\indanevi	\indaneh	\indanehi
6,6	\decalinev		\decalineh	
	\decalinevb	\decalinevt		
6,6,6	\hanthracenev	\hphenanthrenev		

Table 7. Specific Skeletons for N-Containing Heterocycles (\SpecSkel)

N	N-containing heterocyclic skeleton			
1	\pyridinev	\pyridinevi	\pyridineh	\pyridinehi
2,4	\pyrazinev		\pyrazineh	
1,3	\pyrimidinev	\pyrimidinevi	\pyrimidineh	\pyrimidinehi
1,2	\pyridazinev	\pyridazinevi	\pyridazineh	\pyridazinehi
1,3,5	\triazinev	\triazinevi	\triazineh	\triazinehi

Table 8. Carbocyclic Aromatic Skeletons (\CaromSkel)

ring size	carbocyclic aromatic skeleton		
6	\bzdrv		\bzdrh
6,6	\naphdrv		\naphdrh
	\naphdrvb	\naphdrvt	
	\tetralinev		\tetralineh
	\tetralinevb	\tetralinevt	
6,6,6	\anthracenev	\phenanthrenev	


```
\sixheterov      \cyclohexanev    \sixheterov      \pyridinev
[a]{}{1==F}      [a]{1==F}        [ace]{1==N}{2==F}  [ace]{2==F}
```

Figure 4. Synonymous XyM notations for derivatives from generic and specific skeletons.


```
\decaheterov     \decalinev       \naphdrv
[acfhk]{}{1==F}  [acfhk]{1==F}    {1==F}
```

Figure 5. Synonymous XyM notations for derivatives from generic, specific, and carbocyclic aromatic skeletons.

a linking unit, while \lyl is the left-hand counterpart. These intervening skeletons take two arguments: \ryl(LINK)-{GROUP} and \lyl(LINK){GROUP}. The first argument LINK in the parentheses indicates an intervening unit with an integer (0−8) showing the slope of a left or right incidental bond (L0 to L8 for \ryl and R0 to R8 for \lyl, as shown in Figure 1). The former unit is separated from the latter integer with an intermediate delimiter (==). The top-right example of Table 9 has 5==N=N as a LINK.

The second argument GROUP of \ryl or \lyl is a substituent produced by a yl function. The top-right example of Table 9 has {3==\sixheterov{}{5==(yl)}} as a {GROUP}, where the number 3 before a delimiter (==) indicates the slope of a right or left incidental bond, being selected from integers

Table 9. Extended Derivation Processes for XyM Notation

derivation process	derivative family produced from \sixheterov	
substitution		
	\sixheterov{}{2==\sixheterov{}{5==(yl)}}	\sixheterov{}{2==\ryl(5==N=N){3==\sixheterov{}{5==(yl)}}}
atom		
	\sixheterov{2s==\sixheterov{}{5==(yl)}}{}	\sixheterov{2h==\sixheterov{5==$_{+}$N}{5==(yl)}}{}
bond		
	\sixheterov[[b\sixfusev{}{}{E}]]{}{}	\sixheterov[[b\threefuseh{}{}{B}]]{}{}
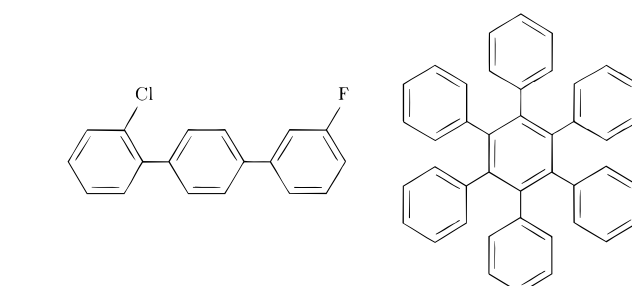
(left)
```
\cyclohexaneh[ace]{%
1==\cyclohexaneh[ace]{4==(yl);3==Cl};%
4==\cyclohexaneh[ace]{1==(yl);3==F}}
```
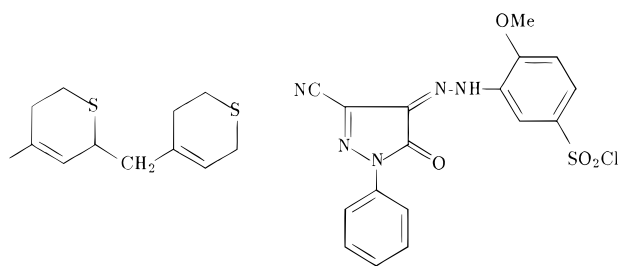
(right)
```
\cyclohexanev[ace]{1==\cyclohexanev[ace]{4==(yl)};%
2==\cyclohexanev[ace]{5==(yl)};3==\cyclohexanev[ace]{6==(yl)};%
4==\cyclohexanev[ace]{1==(yl)};5==\cyclohexanev[ace]{2==(yl)};%
6==\cyclohexanev[ace]{3==(yl)}}
```

Figure 6. yl functions multiply placed in a SUBSLIST.

0 to 8 (R0 to R8 for \ryl and L0 to L8 for \lyl, as shown in Figure 1). Other examples are listed in Figure 7. Note that $_{2}$ is used to represent a subscript and that \bzdrh is the abbreviated skeleton of \cyclohexaneh[bdf].

3.3. Atom Derivation for Spiro Rings. A substituent generated by the yl function can be written in the ATOM-LIST of another skeleton in order to yield a spiro ring system. This is an extension of the atom derivation, as shown in Table 9. The left and right formulas of the atom-derivation row show the absence and the presence of a heteroatom at the spiro position (spiro[5.5]undecane and its aza analogue). Note
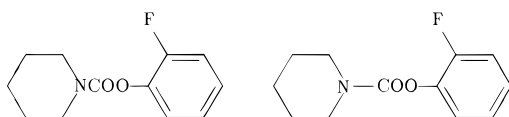
XYM NOTATION FOR ELECTRONIC COMMUNICATION

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **909**

(left)

```
\sixheterov[d]{2==S}{5==\null;%
3==\ryl(3==CH$_{2}$){3==\sixheterov[d]{2==S}{5==(yl)}}}
```

(right)

```
\fiveheterov[d]{1==N;5==N}{4==NC;1==\bzdrv{1==(yl)};2D==O;%
3D==\ryl(5==N-NH){4==\bzdrh{1==(yl);2==OMe;5==SO$_{2}$Cl}}}
```

Figure 7. Linking by \ryl.


```
\sixheterohi{1h==\ryl(4==NCOO){4==\bzdrh{1==(yl);2==F}}}{}     (left)
\sixheterohi{1==N}{1==\ryl(4==COO){4==\bzdrh{1==(yl);2==F}}} (right)
```

Figure 8. ATOMLIST vs SUBSLIST for linking by \ryl.

that the atom modifier s in the notation 2s==\sixheterov-{}{5==(yl)} represents no heteroatom at the spiro position. When a heteroatom is present at the spiro position, the atom modifier h is used in place of s, as found in the notation 2h==\sixheterov{5==$_{+}$N}{5==(yl)}.

It should be emphasized here that spiro[5.5]undecane (named in terms of the IUPAC Rules A-41.1 and A-41.2) is alternatively regarded as "cyclohexana"−cyclohexane so as to give the XyM notation. This methodology is related to the IUPAC replacement nomenclature, e.g., oxacyclohexane (more formally, oxane or tetrahydropyran). IUPAC Rule A-42.1 gives a more sound basis to the XyM notation, since the spiro compound is named cyclohexanespirocyclohexane, where the "cyclohexanespiro" shows the replacement of a carbon atom in a cyclohexane by another cyclohexane ring. Note that the cyclohexanespiro moiety corresponds to the notation 2s==\sixheterov{}{5==(yl)}, which is generated by the yl function.

A substituent generated by \ryl can be also used in an ATOMLIST. The left formula of Figure 8 illustrates such usage, while the right formula indicates the usage of \ryl in a SUBSLIST. This pair of examples shows the flexibility of the XyM notation system so that structural formulas equivalent chemically can be generated by the atom derivation as well as by the substitution one.

3.4. Bond Derivation for Ring Fusion. A fusing skeleton \FuseSkel is defined to treat ring fusion \FuseSkel-[BONDLIST]{ATOMLIST}{SUBSLIST}{FUSE} as listed in Table 10. Each \FuseSkel has the same arguments as those of \GenSkel except for the argument {FUSE}, which designates a bond to be fused after deletion. The argument FUSE is a lowercase alphabet letter or its uppercase counterpart, each of which is a bond specifier representing one bond to be omitted and either terminal of the bond to

Table 10. Fusing Skeletons

ring size	fusing skeleton			
3	\threefusev	\threefusevi	\threefuseh	\threefusehi
4	\fourfuse			
5	\fivefusev	\fivefusevi	\fivefuseh	\fivefusehi
6	\sixfusev	\sixfusevi	\sixfuseh	\sixfusehi
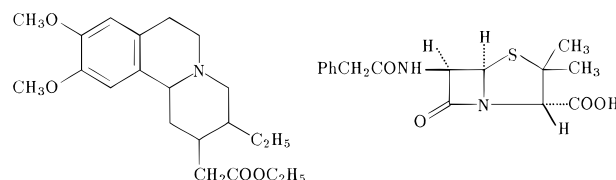

```
\decaheterov[fhk{c\sixfusev[]{1==\null}%
{3==C$_2$H$_5$;4==CH$_2$COOC$_2$H$_5$}{F}}]
{3==N}{6==CH$_3$O;7==CH$_3$O}                    (left)

\fiveheterovi[{d\fourfuse{2==\null}
{1D==O;4Su==PhCH$_{2}$CONH;4Sd==H}{b}}]{1==S;4==N}
{2Sa==CH$_{3}$$;2Sb==CH$_{3}$$;3SA==COOH;3SB==H;5GA==H}     (right)
```

Figure 9. Heteroatoms on fused bonds.

be taken into consideration. Thus, a lowercase character (e.g., a) represents the younger terminal of the omitted bond (e.g., bond a) as a reference terminal. The corresponding uppercase character (e.g., A) designates the other terminal of the bond (e.g., a) to be omitted. The word "older" or "younger" is concerned with the order of numbering vertices. As for a six-membered ring, the numbering 1−2−3−4−5−6−1 shows that terminal 1 (designated by the letter a) of bond a (1−2) is youger, while terminal 2 (designated by the letter A) of bond a is older. It should be noted that terminal 6 of bond f (6−1) is younger, while terminal 1 of bond f is older.

Each of the resulting fusing units is paired with an alphabet letter (lowercase or uppercase for indicating a reference terminal) and, after bracing, placed in the BONDLIST of an outer skeleton. The left formula in the bond-derivation row of Table 9 contains the pair of b (in a mother skeleton) and E (in a fusing unit), as found in the notation {b\sixfusev-{}{}{E}}. This means that the younger teminal in the mother skeleton (b) is placed on the older terminal in the fusing unit (E) so that bond b of the mother skeleton and bond e of the fusing unit are superimposed. The right formula in the bond-derivation row of Table 9 indicates the fusion of a three-membered unit to a six-membered skeleton. Figure 9 illustrates ring fusions in which a heteroatom is incorporated into the fused position.

A skeleton \FuseSkel for ring fusion is capable of accommodating another \FuseSkel skeleton in a nested fashion. By this technique, the carbazole structure, which is generated by attaching a six-membered ring (\sixfusev[ac]-{}{}{e}) to bond b of an indole nucleus, can take a further fused ring so as to produce the structural formula of 7*H*-pyrazino[2,3-*c*]carbazole, as shown in the code

\nonaheterov[begj{b\sixfusev[ac{a\sixfusev[bf]

{6==N;3==N}{}{}{D}}]{}{}{e}}]{1==N}{1==H}

Thus, the construction follows the order of 6,5 ← 6 ← 6.

The structural formula of pyrido[1′,2′:1,2]imidazo[4,5-*b*]-quinoxaline can be obtained by the XyM notation of the order

of 6,6 ← 5 ← 6, as found in the code

\decaheterov[acegi{b\fivefusev[a{b\sixfusev[ac]

{6==\null}{}{e}]{1==N;3==N}{}{d}}]{1==N;

4==N}{}

The IUPAC name corresponds to a quinoxaline with a fused five-membered ring (an imidazo moiety) which is in turn fused by a six-membered ring (a pyrido moiety). The order of constructing the IUPAC name is realized in the code with nested ring fusion. Note that the indicators 1′,2′ and 1,2 of the locant [1′,2′:1,2] in the IUPAC name correspond respectively to the bond specifiers E and b appearing in the code {b\sixfusev[ac]{6==\null}{}{E}}. On the other hand, the indicators, 4,5 and b of the locant [4,5-*b*] are respectively associated with the specifiers d and b appearing in the code {b\fivefusev[...]{1==N;3==N}{}{d}}.

4. EXTENSION FOR STEREOCHEMICAL INFORMATION AND BOND DELETION

4.1. Stereochemistry of Skeletal Bonds. In addition to the arguments described in the preceding sections, the full formats of \GenSkel and \FuseSkel involve a top optional argument SKBONDLIST for treating stereochemical information as well as an end optional argument OMITLIST for treating a bond-deleted skeleton:
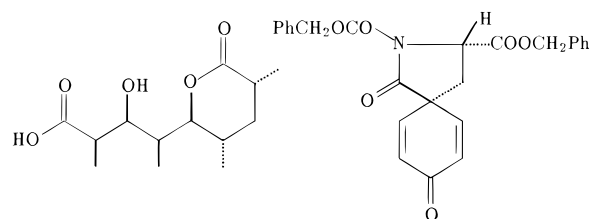
\GenSkel(SKBONDLIST)[BONDLIST]{ATOMLIST}
{SUBSLIST}[OMITLIST]

\FuseSkel(SKBONDLIST)[BONDLIST]{ATOMLIST}
{SUBSLIST}[OMITLIST]

The argument SKBONDLIST contains pairs of two letters in braces, where each pair consists of a bond specifier (a lowercase letter) and an uppercase letter (A or B). The letter A represents an α (downward) bond, while B represents a β (upward) bond. For example, an SKBONDLIST ({aA}{cB}) shows that bond a is an α bond in a dotted form and that bond c is a β bond in a boldfaced form. The argument OMITLIST is a list of bond specifiers, each of which designates a bond to be deleted. As a matter of course, SKBONDLIST and OMITLIST take no common bond specifiers.

The first example of Figure 10 shows that the \sixheterov skeleton takes an optional SKBONDLIST ({eB}) which gives a boldfaced bond at e in the resulting tetrahydropyran ring. This notation is based on the replacement technique, in which the BONDLIST of the \sixheterov skeleton is used for specifying the side chain. The second structure of Figure 10 shows the use of SKBONDLIST for giving a spiro ring. The third example of Figure 10 contains two pairs of stereochemical information in the SKBONDLIST of the \sixheterov skeleton. It should be added here that the XyM notation system permits Chemical Abstracts abbreviations such as Me (for a methyl group) and Ph (for a phenyl group) as well as conventions such as OH (for a hydroxyl group) and COOH (for a carboxyl group), since it aims at drawing or displaying structural formulas in various styles required for electronic communication.

4.2. Deletion of Skeletal Bonds. The OMITLIST argument is used to generate a large ring. The first formula of

\sixheterov({eB}){6==O;5s==\pentamethylenei{}
{1W==HO;1D==O;2B==\null;3B==OH;4B==\null;5==(yl)}}
{1D==O;2A==\null;4A==\null}   (left)

\sixheterov[be]{%
1s==\fiveheterov({aA}{eB}){4==N}%
{4==PhCH$_{2}$OCO;3SB==H;3SA==COOCH$_{2}$Ph;%
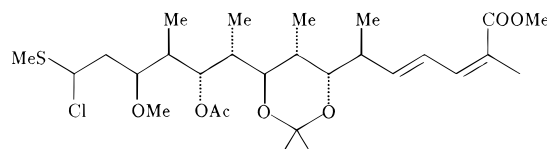5D==O;1==(yl)}}{4D==O}   (right)

Figure 10. Stereochemistry of skeletal bonds.
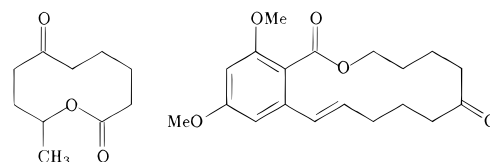
\decaheterov{9==O}{4D==O;8D==O;5==CH$_{3}$$}[k]   (left)

\decaheterov[cegi{b\sixfusev[{b\sixfusev{}{3D==O}{E}]
{6==O}{}{E}[b]}]{2==\null}{6==MeO;8==OMe;1D==O}[b]   (right)
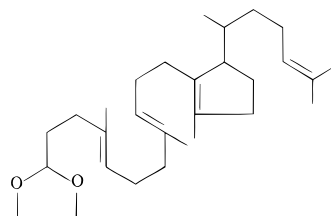
\decaheterov[k{f\fivefusevi{2==\null;5==O}{}{A}}%
{a\sixfusev[d{b\fivefusevi[d{a\sixfusev{%
3s==\trimethylenei[a]{}{1==(yl);2==\null}}
{6==\null}{D}[c]}]{}{}{D}}]{}{3G==\null}{D}[c]}]
{5==O}{{10}Sb==\null;2G==\null}[ej]

Figure 11. Deletion of skeletal bonds.

Figure 11 illustrates a simple case in which one bond is deleted to generate a 10-membered ring. The second formula of Figure 11 contains one \decaheterov skeleton with bond deletion as well as two \sixfusev units with bond deletion so as to give a 14-membered ring. The third formula, which is an intermediate for synthesizing steroid skeletons, can also been generated by the XyM notation due to this technique.

5. DISCUSSION

5.1. Nonunique Nature of the XyM Notation System. A notation or nomenclature for communication should have a nonunique nature. For example, the IUPAC nomenclature for organic compounds has such nonunique nature. Thus,

XᴙM Notation for Electronic Communication

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **911**

Figure 12. Nonunique nature of XyM notations and IUPAC names.

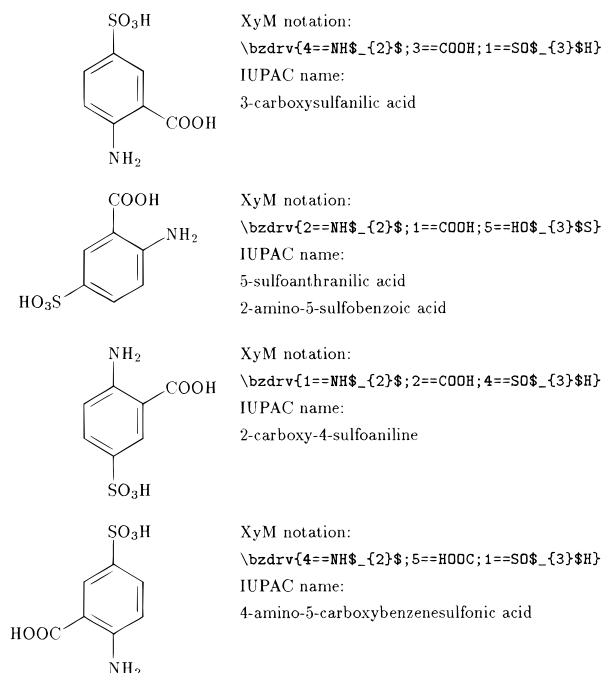3-carboxysulfanilic acid, 5-sulfoanthranilic acid, 2-amino-5-sulfobenzoic acid, 2-carboxy-4-sulfoaniline, and 4-amino-5-carboxybenzenesulfonic acid are synonymous IUPAC names for designating a sole compound.[6] The nonunique nature is convenient for communication, since discussions on a series of sulfanilic acids (4-aminobenzenesulfonic acid) requires the name 3-carboxysulfanilic acid, those on a series of anthranilic acids (2-aminobenzoic acid) take into account the name 5-sulfoanthranilic acid, and so on. These synonyms are translated into such XyM notations as listed in Figure 12, if each of the most predominant substituents is placed at the top of a benzene nucleus. It should be emphasized that the SUBSLISTS of the XyM notations in Figure 12 are different only in their locant numbers. This fact is an advantage of the XyM notation system over the IUPAC nomenclature, since a series of derivatives can be described in a more systematic manner. For example, a series of derivatives of sulfanilic acid can be generated simply by adding a substituent (e.g., 3==COOH for 3-carboxysulfanilic acid) to the SUBSLIST of the XyM notation of sulfanilic acid, i.e., \bzdrv{4==NH$_{2}$;1==SO$_{3}$H}.

The second example of Figure 12 shows that the XyM notations for 5-sulfoanthranilic acid and for 2-amino-5-sulfobenzoic acid take the same format, which is based on the XyM notation for anthranilic acid and 2-aminobenzoic acid, i.e., \bzdrv{2==NH$_{2}$;1==COOH}. The latter IUPAC name is in turn related to the XyM notation, \bzdrv{1==COOH}, for benzoic acid. Thus, the relationship between derivatives is indicated by the XyM notation system more directly than the IUPAC nomanclature system. This nature of the XyM notation stems from the introduction of the concept "substitution derivation" ascribed to SUBSLISTs.

5.2. Interchangeable View of Skeletons and Substituents. An interchangeable view of skeletons and substituents is allowed in the XyM notation system. Fluorobenzene generated by the skeleton \cyclohexaneh[ace]{3==F} is converted into a substituent (i.e., 3-fluorophenyl) by adding the code (yl) so as to give \cyclohexaneh[ace]{1==(yl);

3==F}. Then, the resulting code is added to the SUBSLIST of another skeleton \cyclohexaneh to give the code \cyclohexaneh[ace]{1==Cl;4==\cyclohexaneh[ace]{1==(yl); 3==F}}. This construction corresponds to the IUPAC name 1-chloro-4-(3-fluorophenyl)benzene, which is given by regarding one benzene nucleus as a mother skeleton. When the other benzene nucleus is regarded as a mother skeleton, we can obtain another XyM notation giving the same structure, \cyclohexaneh[ace]{1==\cyclohexaneh[ace]{1==Cl; 4==(yl)};3==F}, which corresponds to the name of the inverse viewpoint, i.e., 1-fluoro-3-(4-chlorophenyl)benzene.

The structural formula of 1-chloro-4-morphorinobenzene can be drawn in two different ways. In the notation \cyclohexaneh[ace]{1==Cl;4==\sixheteroh[]{1==N; 4==O}{1==(yl)}}, the morphorino group is regarded as a substituent, as the name 1-chloro-4-morphorinobenzene indicates. On the other hand, the chlorophenyl group is considered to be a substituent in the notation

\sixheteroh[]{1==N;4==O}{1==\cyclohexaneh[ace]
{1==Cl;4==(yl)}}

so as to correspond to the name *N*-(4-chlorophenyl)-morphorine.

5.3. Vertical and Horizontal Formulas. The electronic communication of structural formulas provides other features than the traditional communication via IUPAC nomenclature, etc. If a structural formula is generated by such an IUPAC name as 1-fluorobenzene, the resulting diagram may be drawn in the vertical, horizontal, or other direction. This feature of the IUPAC name assumes that the structure is communicated in an abstract manner (i.e., without layout data) or by the attachment of a drawn diagram. On the other hand, the electronic communication of structural formulas is by itself used to generate a diagram in print or on a display. Hence, it requires a method for the reproduction of layout data for the vertical, horizontal, or other direction. The XyM notation system takes into account this requirement, giving \bzdrv{1==F} for the vertical drawing and \bzdrh{1==F} for the horizontal drawing. The suffixes v and h designate the directions of the resulting diagrams. Further specification is necessary for fused rings, as illustrated in Figure 3.

The differentiation of skeletons by suffixes provides us with additional merits, as illustrated in Figure 13. The upper-left formula shows that fusing units such as \fivefusev can be multiply nested in itself and in other types of fusing units. When all of the skeletons in the upper-left are changed into their inverse counterparts (\decaheterovi to \decaheterov; \fivefusev to \fivefusevi; and \sixfusev to \sixfusevi), the XyM notation is transformed into another format, giving the formula of the vertically inverse type (the upper-right formula of Figure 13).

The multiply nested example for drawing a structure of the vertical type can be changed into the corresponding horizontal type (the bottom-left formula in Figure 13) if all of the skeletons are changed into horizontal types (\decaheterovi to \decaheterohi; \fivefusev to \fivefuseh; and \sixfusev to \sixfuseh). When all the skeletons in the above notation are changed into their inverse counterparts (\decaheterohi to \decaheteroh; \fivefuseh to \fivefusehi; and \sixfuseh to \sixfusehi), the notation is transformed into another one (the bottom-right formula in Figure 13).
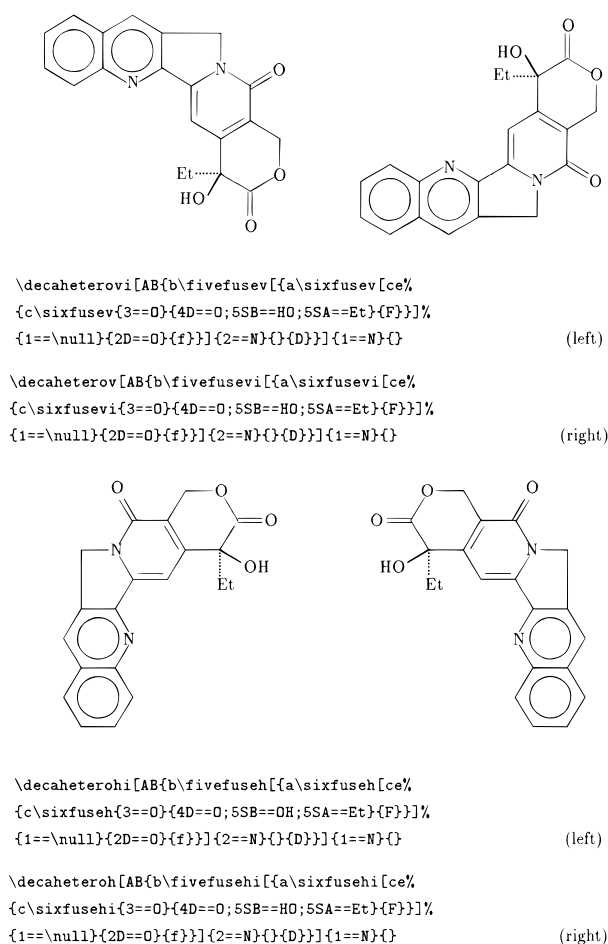
```
\decaheterovi[AB{b\fivefusev[{a\sixfusev[ce%
{c\sixfusev{3==O}{4D==O;5SB==HO;5SA==Et}{F}}]%
{1==\null}{2D==O}{f}}]{2==N}{}{D}}]{1==N}{}        (left)
```

```
\decaheterov[AB{b\fivefusevi[{a\sixfusevi[ce%
{c\sixfusevi{3==O}{4D==O;5SB==HO;5SA==Et}{F}}]%
{1==\null}{2D==O}{f}}]{2==N}{}{D}}]{1==N}{}        (right)
```



```
\decaheterohi[AB{b\fivefuseh[{a\sixfuseh[ce%
{c\sixfuseh{3==O}{4D==O;5SB==OH;5SA==Et}{F}}]%
{1==\null}{2D==O}{f}}]{2==N}{}{D}}]{1==N}{}        (left)
```

```
\decaheteroh[AB{b\fivefusehi[{a\sixfusehi[ce%
{c\sixfusehi{3==O}{4D==O;5SB==HO;5SA==Et}{F}}]%
{1==\null}{2D==O}{f}}]{2==N}{}{D}}]{1==N}{}        (right)
```

Figure 13. Equivalent formulas in different directions.

It is worthwhile to compare the XyM notation system and the CML. In the XyM notation system, equivalent formulas are differentiated in terms of suffixes such as v and h, as found in Figure 13. On the other hand, the CML[61] represents structure data as a kind of connection table, where a structure is drawn by means of *x* and *y* coordinates. This means that such equivalent formulas have distinct sets of *x* and *y* coordinates, which are separately generated by means of an appropriate software for drawing chemical structures.

5.4. The yl Function for Substitution and Replacement. A substituent generated by the yl function is capable of working as a functional group for the substitution derivation as well as a spiro moiety for the atom derivation. For example, see \sixheterov{}{5==(yl)} for the substitution derivation and for the atom derivation in Table 9. This feature of the XyM notation is different from the methodology of the IUPAC nomenclature system. Note that a substituent generated by the yl function is considered to be a *structure without a bond*.

When the substituent is placed in the SUBSLIST of a skeleton, a linking bond is given by the skeleton in the XyM notation system. On the other hand, the IUPAC nomenclature system takes a different methodology so that a linking bond belongs to a substituent, as found in its substitution nomenclature (e.g., cyclohexylcyclohexane). Thus, the group name "cyclohexyl" represents a univalent moiety.

In the XyM notation system, the same substituent can be used in the ATOMLIST of a skeleton, as shown in the atom derivation in Table 9. It is regarded as a replacement unit so

Table 11. Package Files of XyMTeX

package	included functions
aliphat.sty	macros for drawing aliphatic compds
carom.sty	macros for drawing vertical- and horizontal-type carbocyclic compds
lowcycle.sty	macros for drawing five-or-less-membered carbocycles
ccycle.sty	macros for drawing bicyclic compounds, etc.
hetarom.sty	macros for drawing vertical-type heterocyclic compds
hetaromh.sty	macros for drawing horizontal-type heterocyclic compds
hcycle.sty	macros for drawing pyranose and furanose derivatives
chemstr.sty	basic commands for atom and bond typesetting
locant.sty	commands for printing locant numbers
polymers.sty	commands for drawing polymers
fusering.sty	commands for drawing units for ring fusion
methylen.sty	commands for drawing zigzag polymethylene chains
xymtex.sty	package for calling all package files
chemist.sty	commands for using "chem" version and chemical environments

as to produce a spiro ring. This methodology is parallel to the replacement nomenclature (e.g., IUPAC Rule A-42.1), which is one of the predominant methods for naming spiro compounds. For example, cyclohexanespiro in the name cyclohexanespirocyclohexane shows the replacement of a carbon atom in a cyclohexane by another cyclohexane ring, since the moiety cyclohexanespiro is considered to represent a structure without a bond.

It should be emphasized that the concept of "substituent without a bond" has been made possible by considering SUBSLISTs and ATOMLISTs as the distinct lists of descriptors. In contrast, the IUPAC nomenclature system adopts differentiation methods by suffixes, where it inevitably requires "substituents with a linking bond" in its substitution nomenclature.

6. IMPLEMENTATION

The XyMTeX system (version 1.01)[58,59] has not supported the XyM notation described in this paper. Hence, it has required layout data to combine two or more structures. For example, biphenyl has been generated by the code containing layout data, \bzdrh{4==\kern-25pt\lower37pt\hbox{\bzdrh-{}}}, in which the commands \kern (for horizontal adjustment), \lower (for vertical adjustment), and \hbox (for generation of a typesetting box) are TeX commands to adjust the *x,y* coordinates of the substitution site.[66] On the other hand, the XyM notation for generating an equivalent structure, \bzdrh{4==\bzdrh{}}, has no layout data, which should be added automatically during the processing of the XyM notation. In other words, the extended functions of substitution and atom and bond derivations should be supported for the implementation of the XyM notation.

The XyM notation system has been implemented as an enhanced version of the XyMTeX system (version 2.00),[67,68] in which each XyM notation is regarded as a XyMTeX command. The enhanced XyMTeX system, which consists of package files listed in Table 11, has been designed to support the extended functions described in this paper. The package file chemstr.sty is the basic file that is automatically read within any other package file of XyMTeX. It contains macros for internal use, e.g., common commands for bond and atom setting. The other package files contain macros for users. These files are designed to work as packages for LaTeX2e.[69] The use of xymtex.sty calling all package files

XyM Notation for Electronic Communication

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **913**

may sometimes cause the "TeX capacity exceeded" error. In this case, you should call necessary packages distinctly by using the \usepackage command.

The structural formulas involved in all the figures and tables of this paper have been generated by XyMTeX (version 2.00), where the XyM notations cited there are input and interpreted as XyMTeX commands.

7. CONCLUSION

The XyM notation system is proposed as a linear notation for the electronic communication of structural formulas. A derivative family is considered to discuss derivatives coming from a skeleton, where three derivation processes are taken into account: substitution derivation for placing large substituents, atom derivation for spiro fusion, and bond derivation for ring fusion. The derivations are realized by using the arguments of the skeleton (SUBSLIST, ATOMLIST, and BONDLIST). Stereochemical and ring-opening information are discussed by means of additional arguments (SKBONDLIST and OMITLIST). The XyM notation system is implemented as a LaTeX2e application.

REFERENCES AND NOTES

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

(2) Warr, W. A., Ed. *Chemical Structure Information Systems. Interfaces, Communication, and Standards*; ACS Symposium Series 400; American Chemical Society: Washington, DC, 1989.

(3) Bawden, D., Mitchell, E. M., Eds. *Chemical Information Systems. Beyond the Structure Diagram*; Ellis Horwood: Chichester, 1990.

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

(5) Ono, S., Ed. *Computer Chemistry*; Maruzen: Tokyo, 1988.

(6) Hirayama, K. In *Joho kagaku (Information Chemistry)*; Chemical Society of Japan, Ed.; Gakkai Shippan Center: Tokyo, 1979; Kagaku Sosetu Vol. 18; pp 9−28.

(7) Kudo, Y. In *Joho kagaku (Information Chemistry)*; Chemical Society of Japan, Ed.; Gakkai Shippan Center: Tokyo, 1979; Kagaku Sosetu Vol. 18; pp 63−90.

(8) Hanai, S. In *Computer Chemistry*; Ono, S., Ed.; Maruzen: Tokyo, 1988; pp 57−98.

(9) Wiswesser, W. J. Historic development of chemical notations. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258−263.

(10) Davis, C. H.; Rush, J. E. *Information Retrieval and Documentation in Chemistry*; Greenwood: Westport, 1974.

(11) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood: Chichester, 1985.

(12) National Academy of Sciences−*National Research Council Survey of Chemical Notation Systems*; National Academy of Sciences: Washington, DC, 1964.

(13) *National Academy of Sciences Chemical Structure Information Handling. A Review of the Literature 1962−1968*; National Academy of Sciences: Washington, DC, 1969.

(14) *The Commission on Codification, Ciphering and Punched Card Techniques of the IUPAC Rules for IUPAC Notation for Orgaic Compounds*; John Wiley & Sons: New York, 1961.

(15) Dyson, G. M. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, 1974; pp 130−155.

(16) Smith, E. G., Baker, P. A., Eds. *The Wiswesser Line-Formula Chemical Notation (WLN)*, 3rd ed.; Chemical Information Management: Cherry Hill, 1976.

(17) Hayward, H. W.; Sneed, H. M. S.; Turnipseed, J. H.; Tauber, S. J. Some experience with the Hayward linear notation system. *J. Chem. Doc.* **1965**, *5*, 183−189.

(18) Skolnik, H. A new linear notation system based on combinations of carbons and hydrogen. *J. Heterocyclic Chem.* **1969**, *6*, 689−695.

(19) 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.

(20) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(21) Weininger, D. SMILES. 3. Depict. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237−243.

(22) Wilcox, C. S.; Levinson, R. A. In *Artificial Intelligence Applications in Chemistry*; ACS Symposium Series 306; Pierree, T. H., Hohne, B. A., Eds.; American Chemical Society: Washington, DC, 1986; p 228.

(23) Ghoshal, N. Yet nother linear notation scheme for organic compounds. 1. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 308−311.

(24) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(25) Mooers, C. N. Ciphering structurl formulas−the Zatopleg system. *Zator Tech. Bull.* **1951**, *59*.

(26) Ray, L. C.; Kirsch, R. A. Finding chemical records by digital computers. *Science* **1957**, *126*, 814−819.

(27) Gluck, D. J. A chemical structure storage and search system devloped by at DuPont. *J. Chem. Doc.* **1962**, *5*, 43−51.

(28) Spialter, L. The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): A New computer-oriented chemical nomenclature. *J. Am. Chem. Soc.* **1963**, *85*, 2012−2013.

(29) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurt, A. K. I.; Greier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(30) Kasparek, S. V. *Computer Graphics and Chemical Structures*; John Wiley & Sons: New York, 1990.

(31) Morgan, H. L. The generation of a unique machine description for chemical structures−a technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(32) Petrarca, A. E.; Lynch, M. F.; Ruch, J. E. A method for generating unique computer structural representations of stereoisomers. *J. Chem. Doc.* **1967**, *7*, 154−165.

(33) Shelley, C. A.; Munk, M. E. An approach to the assignment of canonical connection tables and topological symmetry perception. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247−250.

(34) Herndon, W. C.; Leonard, J. E. Canonical numbering, sterochemical descriptors, and unique linear notaions for polyhedral clusters. *Inorg. Chem.* **1983**, *22*, 554−557.

(35) Schubert, W.; Ugi, I. Constitutional symmetry and unique descriptors of molecules. *J. Am. Chem. Soc.* **1978**, *100*, 37−41.

(36) Randić, M. On unique numbering of atoms and unique codes for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105−108.

(37) Hendrickson, J. B.; Toczko, A. G. Unique numbering and cataloguing of molecular structures. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171−177.

(38) Jochum, C.; Gasteiger, J. On the misinterpretation of our algorithm for the perception of consitutional symmetry. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 49−50.

(39) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.

(40) Uchino, M. Algorithms for unique and unambiguous coding and symmetry perception of molecular structure diagarams. 5. Unique coding by the method of "orbit graphs". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 201−206.

(41) Fujita, S. Canonical numbering and coding of imaginary transition structures. A novel approach to the linear coding of individual organic reactions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 128−137.

(42) Balaban, A. T.; Mekenyan, O.; Bonchev, D. Unique description of chemical structures based on hierarchicaly ordered extended connectiviyies (HOC procedures). I. Algorithms for finding graph orbits and canonical numbering of atoms. *J. Comput. Chem.* **1985**, *6*, 538−551 and further papers in the series.

(43) Norris, A. C.; Oakley, A. L. In *TeX−Applications, Uses, Methods*; Clark, M., Ed.; Ellis Horwood: Chichester, 1990; pp 207−225.

(44) Norris, A. C.; Oakley, A. L. Electronic publishing and chemical text processing I. A survey of software for two-dimensional chemical structure editing. *Comput. Chem.* **1988**, *12*, 245−251.

(45) Haase, R. T.; O'Kane, K. C. Typesetting chemical structure formulas with the text formatter TeX/LaTeX. *Comput. Chem.* **1987**, *11*, 251−271.

(46) Bachrach, S. M., Ed. *The Internet: A Guide for Chemists*; American Chemical Society: Washington, DC, 1996.

(47) IUPAC Organic Chemistry Division, Commission on Nomenclature of Organic Chemistry. *Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F, and H*, 4th ed.; Pergamon: Oxford, 1979.

(48) IUPAC Organic Chemistry Division, Commission on Nomenclature of Organic Chemistry (III.1). *A Guide to IUPAC Nomenclature of Organic Compounds. Recommendations 1993*; Blackwell: London, 1993.

914 *J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999*

Fujita and Tanaka

(49) Chemical Abstracts. *Index Guide 1992−1996*; Chemical Abstracts Service: Columbus, OH, 1997; Appendix IV.

(50) Hirayama, K. Mechanically interconvertible chemical name, linear notation and structural formula of organic compounds. *Kagaku Ryoiki* **1980**, *34*, 776−783.

(51) Hirayama, K. *The HIRN System. Nomenclature of Organic Chemistry for Man-to-Machine and Man-to-Man Communication. Principles*; Maruzen, Tokyo and Springer-Verlag: Berlin−Heidelberg, 1984.

(52) Cooke-Fox, D. I.; Kirby, G. H.; Rayer, J. D. Computer translation of IUPAC systematic chemical nomenclature. 1. Introduction and background to a grammer-based approach. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 101−105.

(53) Cooke-Fox, D. I.; Kirby, G. H.; Rayer, J. D. Computer translation of IUPAC systematic chemical nomenclature. 2. Development of a formal grammer. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 106−112.

(54) Cooke-Fox, D. I.; Kirby, G. H.; Rayer, J. D. Computer translation of IUPAC systematic chemical nomenclature. 3. Syntax analysis and semantic processing. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 112−118.

(55) Fujita, S. XyMTeX. a macro package for typesetting chemical structural formulas, 1993; On-line Manual for XyMTeX Version 1.00. Available from CTAN (tex-archive/macros/latex209/contrib/xymtex/).

(56) Fujita, S. Typesetting structural formulae with the text formatter TeX/LaTeX. *Comput. Chem.* **1994**, *18*, 109−116.

(57) Fujita, S. XyMTeX for drawing chemical structural formulas. *TUGboats* **1995**, *16*, 80−88.

(58) Fujita, S. XyMTeX. A macro package set for typesetting chemical structural formulas, 1996; On-line Manual for XyMTeX Version 1.01.

(59) Fujita, S. *XyMTeX. Typesetting Chemical Structural Formulas*; Addison-Wesley Japan: Tokyo, 1997.

(60) Ogura, T.; Nishimoto, M.; Hatano, Y. Representation of Mathematical Formula, Chemical Formula and Structral Formula on the WWW Browser; In *The 20th Symposium on Chemical Information and Computer Science, Kumamoto*; 1997; p 21IP15.

(61) Murary-Pust, P.; Rzepa, H. S.; Leach, C. CML−Chemical Markup Language. Presented at the 210th ACS Meeting in Chicago, 1995; Abstract 40 (http://www.chi.ic.ac.uk/cml/).

(62) Americal Chemical Society. CXF, on-line manual, 1995.

(63) ISO/IEC 8824 *Information Tehnology*−Open Systems Interconnection−Specification of Abstract Syntax Notaion One (ASN.1), 1990.

(64) The XyM is an uppercase form of χυμ, where y is used in place of Y in the special layout (XYM). The pronunciation of XyM is recommended to be khým, in which the kh sound may be a Russian kh or, more simply, an English k and the symbol ý is expected to be pronounced like a German ü. Note that the word chemistry stems from the Arabian root alchemy, which is, in turn, considered to come from Greek, χυμείᾱ.

(65) IUPAC Organic Chemistry Division, Commission on Nomenclature of Organic Chemisty Revision of the extended Hantzsch−Widman system of nomenclature for heteromonocycles (Recommendation 1982). *Pure Appl. Chem.* **1983**, *55*, 409−416.

(66) Knuth, D. E. *The TeXbook*; Addison-Wesley: Reading, MA, 1984.

(67) Fujita, S. XyMTeX for typesetting chemical structural formulas. Enhanced functions for version 2.00, 1998; On-line Manual for XyMTeX Version 2.00.

(68) XyMTeX (version 2.00) can be obtained from the author's site, http://www.chem.kit.ac.jp/fujita/fujitas/fujita.html.

(69) Goossens, M.; Mittelbach, F.; Samarin, A. *The LaTeX Companion*; Addison-Wesley: Reading, MA, 1994.

CI990018X