(4) Thomson, L. H., E. Hyde, and F. W. Matthews, "Organic Search and Display using a Connectivity Matrix Derived from Wiswesser Notation," J. CHEM. DOC. 7, 204 (1967).
(5) Lefkovitz, D., R. V. Powers, and H. N. Hill, "CIDS No. 5, Computer Programming for An Experimental Chemical Information and Data System," University of Pennsylvania; produced under Contract DA-18-035-AMC-299(A), Technical Support Directorate, U.S. Army Edgewood Arsenal, Edgewood, Md., 1967.

(6) An MCC branch is any element or combination element symbol with three or more nonhydrogen attachments. Therefore, the symbols L (-(CO)-) and $\alpha X(SO_2, NO_2)$ are not branches.

(7) These screens, without hydrogen denotation, are currently in use by the U.S. Army CIDS and have proved to be highly effective therein.[3]

(8) In logic, this is called a product of sums.

A Line-Formula Notation System for Markush Structures*

HELEN M. S. SNEED, JAMES H. TURNIPSEED, and ROBERT A. TURPIN, JR.
U. S. Patent Office, Department of Commerce, 1406 G St., N. W., Washington, D. C. 20231

Received May 21, 1968

A notation system has been developed in the U. S. Patent Office to handle some Markush forms. The system is presented as a supplement to the existing Hayward Notation System which was developed for specific organic chemical structures. The proposed notation system for organic Markush structures is limited to determinate structures of several isolated Markush forms, including those forms that are restricted in substitution depending on the condition of some other Markush group.

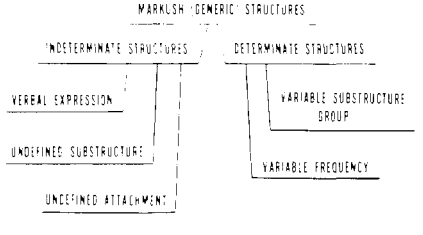
The Line Formula Notation System for Markush Structures is a supplement to a notation system developed for specific organic structures by H. Winston Hayward of the U. S. Patent Office. Efforts were made to avoid some of the rigid unique features of the Hayward system but remain within the over-all framework of the system itself.

Markush structures, as they are referred to in the Patent Office, are generic expressions of chemical structures or structure classes. The expressions can be verbal, structural, or some combination of verbal and structural statements. The Markush expressions which are in terms of structure with all points of attachment defined as *determinate* Markush structures. All other Markush expressions, namely, verbal, combinations of verbal and structural, and structures with undefined attachment, are *indeterminate* Markush structures. For convenience of notation, specific chemical structures may be reduced topologically to a graph in which each atom node (vertex) represents only one atomic structure and each bond (edge) represents only one bond type, and are further defined as a set of atom nodes and bonds such that each bond of the set is connected to two atom nodes. Likewise, determinate Markush structures can be defined in terms of a graph, reducing the Markush group to a node. A determinate Markush structure is then defined as a Markush structure in which all nodes, atom and Markush, are specifically defined by an atomic structure, or group of definite atomic structures and/or atom strings; all bonds are defined by a bond type or group of alternative bond types; and, all points of attachment are explicit.

This investigation will not venture beyond the realm of the determinate Markush structure. It is hoped, however, that the notation system for Markush structures will be modified in the near future to handle some of the indeterminate forms. The determinate Markush forms, as isolated in this report, were extracted from U. S. Patents. For Patent Office purposes, a Markush structure should be copied just as it is disclosed in the original document. This policy reduces the possibility of generating structures which may not be encompassed in the original disclosure.

ISOLATED FORMS OF MARKUSH EXPRESSIONS

To define further the forms in which Markush expressions may be disclosed, the general form may be broken down as follows:

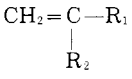


* Presented before the Division of Chemical Literature, Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 4, 1968. This paper reports on work advanced by Patent Office researchers as a part of that agency's continued efforts, in cooperation with the National Bureau of Standards and the National Science Foundation, to solve the problems of retrieval of information from machine-oriented systems containing generic chemical structures.

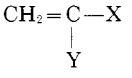
As examples of typical disclosures of indeterminate Markush structures, we have:

A. **Verbal Expression.**
   1. Aliphatic hydrocarbon
   2. Water soluble sulfonates

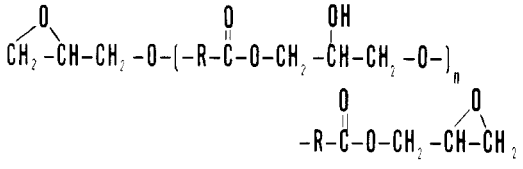
B. **Undefined Substructure.**

1. 

R_1 and R_2 = cycloalkyl, alkyl of 1 to 12 carbon atoms, aryl

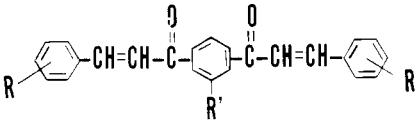
2. 

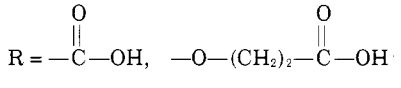
X = halogen, an aliphatic hydrocarbon radical
Y = a monovalent organic radical having the free valence bond attached to a carbon atom and monomers having a terminal methylene group joined to an aliphatic carbon atom through an ethylenic linkage which is in conjugated relationship with another ethylenic linkage, which monomers are capable of forming long-liver polymer free radicals, etc.

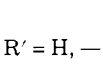
3. 

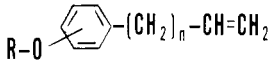
R = an aromatic radical
n = some small whole number

C. **Undefined Attachment.**

1. 





2. 

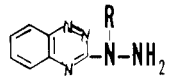
n = 0 to 10
R = straight or branched chain alkyl group containing from 1 to 10 carbon atoms.

Note: The other positions on the benzene ring may be: H, alkyl, cycloalkyl, aryl, halogen, alkoxy, aryloxy and the like.

Of the different types of determinate Markush representations there are two major classes: (1) that class which contains at least one variable substructure group, and (2) that class which contains at least one substructure that occurs a variable number of times (variable frequency). There also may exist combinations of the two cited classes, however, no distinction of the combination is necessary beyond the conclusion that the Markush structure is determinate. For this paper, the two major classes are further divided into subclasses which define in greater detail the function of the respective Markush form being disclosed. The sum of the definitions of the subclasses more clearly defines the class.

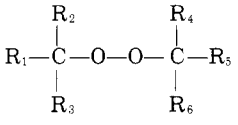
This investigation has isolated five distinct forms of variable substructure groups. They are Markush groups which are:

1. clearly defined by two or more exclusive alternatives (members)



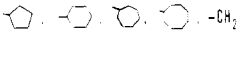
R = H, $-CH_3$, $-C_2H_5$, $CH_3-(CH_2)_2-$,
$(CH_3)_2CH-$, $CH_3-(CH_2)_3-$, $CH_3-(CH_2)_4-$, $CH_3-(CH_2)_5-$,
$CH_3-(CH_2)_6-$, $CH_3-(CH_2)_7-$

2. defined as having the same set of alternatives as a group previously defined, and with no restriction as to the order to substitution;

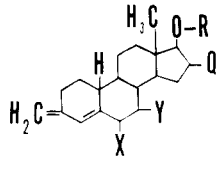


No restriction on substitution

R_1 thru R_6 = $-CH_3$, $-CH_2-CH_3$, $-CH_2-CH_2-CH_3$,
$-CH_2-(CH_2)_2-CH_3$, $-CH_2-(CH_2)_3-CH_3$,
$-CH_2-(CH_2)_4-CH_3$, $-CH_2-(CH_2)_5-CH_3$,
$-CH_2-(CH_2)_6-CH_3$, $-CH_2-(CH_2)_7-CH_3$,
$-CH_2-(CH_2)_8-CH_3$, $-CH_2-(CH_2)_9-CH_3$,
$-CH_2-(CH_2)_{10}-CH_3$,



3. defined as having the same set of alternatives as a group previously defined, with the restriction that the substitution at the current Markush node must not be equal to the substitution at the previously defined Markush node

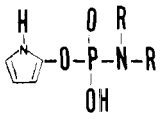


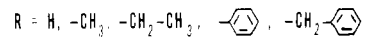
X = H, CH_3-
Y = X (unequal in substitution)

R = H, $CH_3-C(=O)-$, $C_2H_5-C(=O)-$, $CH_3-CH_2-CH_2-C(=O)-$

Q = H, CH_3-, C_2H_5-, $CH_2=CH-$, $CH_2=CH-CH_2-$,
$CH_3-(CH_2)_3-$

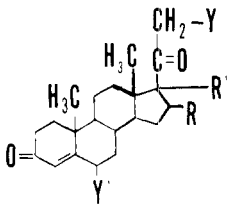
4. defined as having the same members as a group previously defined, with the restriction that the substitution at the current Markush node must be equal to the substitution at the previously defined Markush node

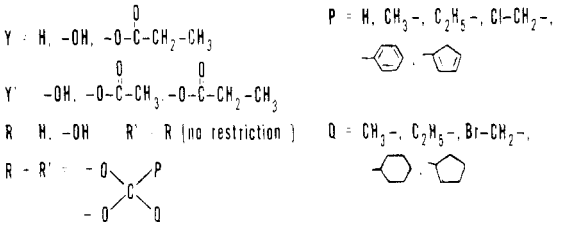




R's must be equal in substitution

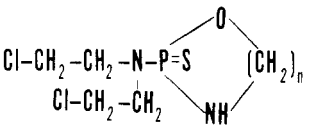
5. defined as representing the combination of two previously defined groups, combined to close a ring.





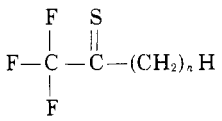
Five subclasses (or forms) of the variable frequency class have been isolated by the present investigation. They are:

1. That form which states a range within definite limits



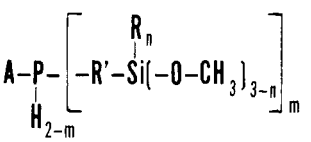
n = 1 to 4

2. That form which states explicitly each value to be substituted



n = 1, 2, 3, 4, 5, 6

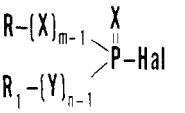
3. that form which states a constant, plus or minus a defined variable



m = 1, 2

n = 0 to 2

4. That form which states a variable plus or minus a constant



m and n = 1, 2

Hal = Br, Cl

R and R_1 = CH_3-, C_2H_5-, CH_3-CH_2-CH_2-, $(CH_3)_2CH$-, CH_3-$(CH_2)_2$-CH_2-

5. that form which defines the current frequency as being the same as one defined previously.

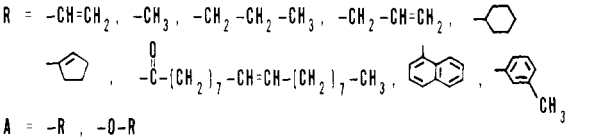
X—$(CH_2)_m$—O—$(CH_2)_n$—X
X = Cl, Br
m = 1, 2
n = m

ORDER OF ENUMERATION AND CIPHERING

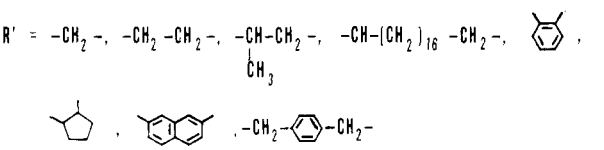
Markush structures are enumerated before ciphering to establish a sense of direction once ciphering is begun. Unlike a specific organic compound, a single Markush structure can have many enumeration patterns. This feature is present because of the flexibility one has in selecting a starting atom. Each potential starting atom will generate one and only one enumeration pattern. The nonunique features of the Markush notation are enhanced by the optional starting atoms, however, once an atom is chosen the enumeration and ciphering are similar to the enumeration and ciphering of specifics. The rules for enumeration of specific organic compounds have been described in the Hayward Notation System.

For the purposes of this notation, each Markush group encountered while enumerating will be treated as a skeletal atom node. After enumerating the skeleton on the Markush structure, each member of the Markush group is treated individually as an isolated substructure and is enumerated beginning with that atom which is connected to the "last cited outside atom" and continued according to the rules of the Hayward Notation System. The last cited outside atom has reference to the last node (atom or Markush) in the structure's skeleton to which the current Markush node is attached.

Markush structures are ciphered by following the established enumeration pattern beginning with the first enumerated atom node. The cipher rules and characters for this notation system are similar to those for specific



A = -R, -O-R

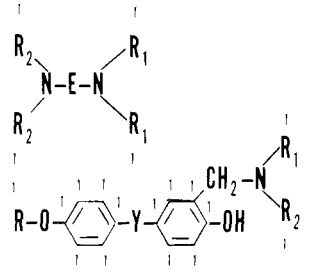


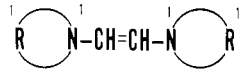
organic compounds with the exception that for Markush structures special rules and symbols are needed to cipher the Markush node and the frequency variables.

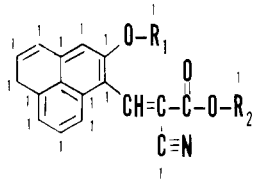
Enumeration is begun on any one of two types of nodes:

a. Terminal skeletal atom (or, Markush) node; or,
b. Peripheral ring atom (or, Markush),

Examples:



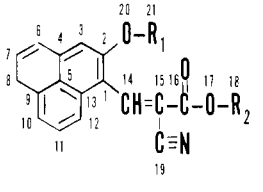




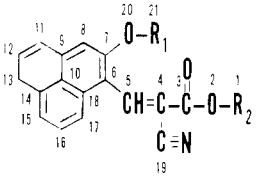
where R, R_1, R_2, E, *and* Y *are Markush nodes.*

Once a starting point on the Markush skeleton has been arbitrarily chosen the sequence of seniority is fixed according to the rules of the Hayward Notation System.

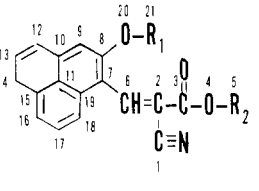
Examples:



or

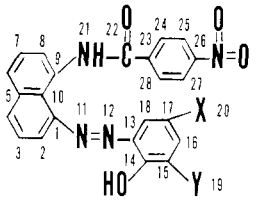


or

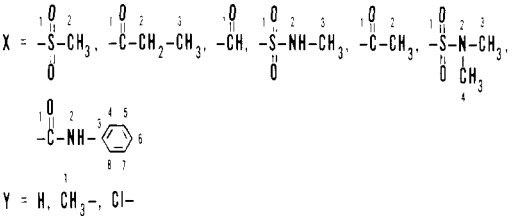


etc.

If given the Markush structure and a possible enumeration pattern for the skeleton,



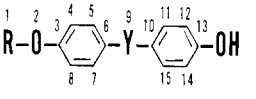
and the Markush groups Y and X are defined as,

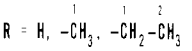


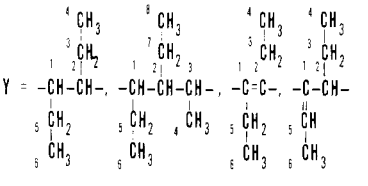
the members are enumerated as shown, beginning with that atom which would be connected to the last cited outside atom if the member were substituted at the Markush node. Note that each member is enumerated individually.

Likewise, where there is more than one connection to a Markush node, the rule of the last-cited outside atom applies.

Example:



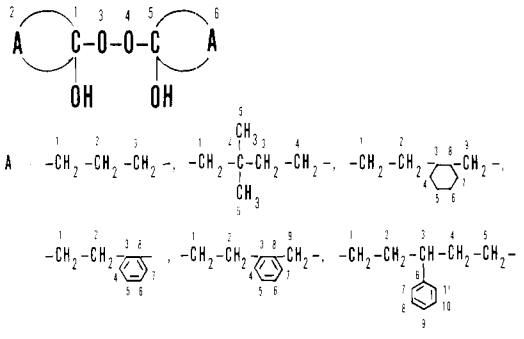




In the second member of Y, the possible enumeration patterns from atom node 2 are symmetrical by the rules of the Hayward Notation System. However, the node enumerated as 3 was determined to be more senior by the rules of the present notation system because it is connected to an outside node. Thus, when enumerating the skeleton of a Markush member which has symmetrical branching paths, that path is senior which has an outside connection.

When a Markush group (node) is embedded in a ring, the enumeration pattern of the members of the group depends entirely upon the positions of the outside connections. Again, the enumeration is begun on that atom node which is connected to the last cited outside atom. The atom nodes that are directly involved in the ring closure must be in the main path of the current member. All other atom nodes (those other than main path nodes) of the current member must then be enumerated according to the rules for enumerating substituents attached to a ring system as disclosed in the Hayward system.

Example:



When there are ring atoms in the path of atoms undergoing ring closure, as in members 3, 4, and 5, the entire ring to which the atoms in question belong is enumerated as being a part of the main path.

Upon completion of the enumeration of a Markush structure, ciphering is started beginning with the lowest enumerated node and is continued according to the rules of the Hayward system. Each Markush group is introduced with the special symbol K followed by a whole number which increases by one for each new Markush group. The K symbol and the numerical value which follows must then be defined by a specific disclosure of the variables that may be substituted at the current node of the structure. The variables (members) are enclosed in pointed brackets (⟨⟩) and each member is separated from the member that follows by a comma (,). Each member of a Markush group is ciphered individually, the members of such group possibly including bond types, and hydrogen (H). Thus, the cipher configurations which correspond to the above mentioned isolated forms of Markush groups are:

1. Kn ⟨A, B, C, etc.⟩
2. Kn ⟨Km⟩
3. Kn ⟨Km≠⟩
4. Kn ⟨Km=⟩
5. $Kn [Kl + Km] ⟨A, B, C, etc.⟩

where A, B, C are symbolic representations of members of Markush groups; n is the numerical value which identifies the current Markush group; l and m are the numerical values which identify Markush groups that were previously defined; and, $ is the special symbol which indicates the combination of two or more Markush nodes to form a ring closure, the nodes being combined are cited within the square brackets ([ ]). This cipher configuration, which represents the combination of two or more Markush groups combining to form a ring, is cited only after all other atom nodes of the Markush structure have been ciphered—i.e., at the end of the cipher.

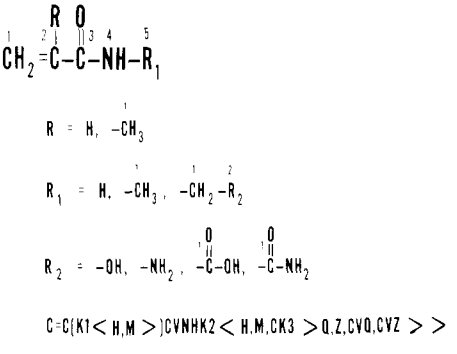
If when ciphering, a substructure group is encountered which is substituted onto an atom node a variable number of times, the cipher of that substructure group is enclosed within parentheses and is followed by the disclosed variable. If the repeating substructure group is embedded in a chain and is not connected to a pivot atom, the cipher of that group is enclosed within braces ({ }) followed by the variable. The pivot atom is that atom which has substituted onto it a substructure group that is repeated around said atom. A varying substructure group embedded in a ring is handled in a different manner, however. The size of such a ring differs according to the value of the variable. Therefore, the ring size must be cited as a variable to introduce the ring and the ring is ciphered in order of its enumeration pattern until the varying substructure group is reached. When encountered, the cipher of that group is enclosed within braces and is followed by the variable. Accordingly, the cipher configurations that correspond to the isolated forms of variable frequencies are:

1. Jx ⟨≥A≤B⟩
2. Jx ⟨A, B, C, etc.⟩
3. Jx ⟨ ± (≥B≤C)⟩; or, Jx ⟨A ± Jw⟩
4. Jx ⟨Jw⟩
5. Jx ⟨z ± A⟩

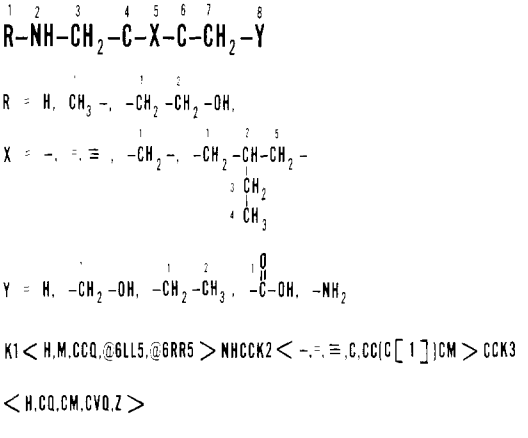
where J is the special symbol for the variable frequency type of Markush disclosure; x is the numerical value which identifies the current variable; A, B, C are whole numbers which identify a variable previously defined; and, z is some variable which may be in any of the above forms.

Having established a methodology for the analysis of the isolated forms of Markush structures, the complete unambiguous ciphers of typical disclosures are as below:

1.

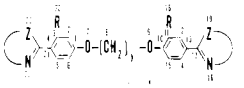


2.



Note the use of the "at" sign (@) before rings as members. This feature is one which conforms to the conventions of the Hayward system where the "at" sign is used to introduce a ring. Note also the use of a connection number in citing the fifth member of K2. Connection numbers are assigned to all connections which are not implicit in the cipher.

3.



Z = $-CH_2-CH_2-CH_2-$; $-CH_2-CH_2-CH_2-CH_2-$

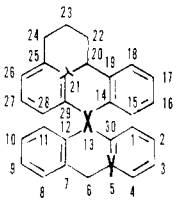
y = 2, 3, 4

R = H, $-O-CH_3$

6R:0,C1)1<2,3,4>0,6RRIK1<H,OM>)RR(,)12<2>5>5,6>L WK2<LLL.

LLLL>!RR:RR(K3<K1>R'.,13<)12>L WK4<K2>RR

4.



X = Si, Ce, Sn, C, Pb, Te, Se, Ge, S

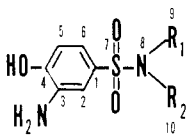
Y = P, N, C, Te, B, In

⅄ = Ga, Yt, Sa, C

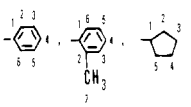
6R4Y<K1<P,N,Y,te,bo,in>>LYR4YX<K2<si,ce,sn,X,pb,te,se,ge,S>>>

YR4YY⅄<K3<ga,yt,sa,⅄>L3YR3Y#Y

5.

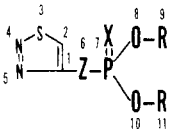


R_1 and R_2 must be unequal in substitution.

R_1 and R_2 = $-CH_3$, $-CH_2-CH_2-CH_3$, 

6R(SV2N(K1<M,CCM,@6RR5,@6RRMR4,@5LL4>)K2<K1≠>)RRZRQRR

6.



R's must be equal in substitution

Z = $-CH_2-O-$, $-CH_2-S-$, $-\overset{O}{\overset{\|}{C}}-$, $-\overset{O}{\overset{\|}{C}}-NH-NH-CH_2-S-$

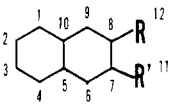
X = O, S

R = $-CH_3$, $-\underset{CH_3}{CH}-CH_3$, $-CH_2-\underset{CH_3}{CH}-CH_3$, $-CH_2-\underset{CH_2}{\underset{|}{CH}}-CH_2-CH_2-CH_2-CH_3$

5L(K1<CO,CS,CV,CVNHNHCS>P(K2<V,=S>)(OK3<M,T,CT,CC(CM)C3M>

)2)N=NSL=

7.



R and R' must be identical in substitution.

R = $-CH_3$, $-CH_2-CH_3$, $-CH_2-CH_2-CH_2-CH_3$

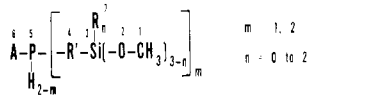
R' = R

R + R' = $-CH_2-CH_2-$, $-CH_2-\underset{CH_3}{CH}-$, $-CH_2-CH_2-CH_2-$, $-\underset{CH_3}{CH}-CH_2-CH_2-$

6L4YLL(K1<M,CM,CCCM>)L(K2<K1=>)LY$K3[K1+K2]<LL,LLM,

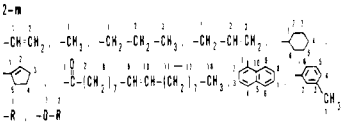
LLL,LMLL>

8.

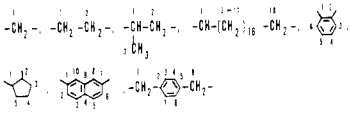


m = 1, 2

n = 0 to 2

R = $-CH-CH_2$, $-CH_3$, $-CH_2-CH_2-CH_3$, $-CH_2-CH-CH_2$, 

A = -R, $-O-R$

R' = $-CH_2-$, $-CH_2-CH_2-$, $-\underset{CH_3}{CH}-CH_2-$, $-CH-(CH_2)_{16}-CH_2-$, 

((MO)I1<3-(≥0≤2)>si(K1<C=C,M,CCM,CC=C,@6LL5,@5L=LL3,

CVC7C=CC7M,@6RR3YR4Y,@6RRRMR3>)K2<C,CC,CMC,CC16C,@6RR[1]

R4,@5LL[1]L3,@6RRRYRRR[1]RYR,C,@6RRRR(C[1]\RR>)I2<

1,2>P[1]M)3<2-I2>K3<K1,OK1>