COMPUTER-AIDED GENERATION OF IUPAC NOMENCLATURES FOR ACYCLIC COMPOUNDS

Hidetsugu ABE, Sae TAKAHASHI and Shin-ichi SASAKI
Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku, Toyohashi 440, Japan

Abstract

The present paper describes a computer-aided IUPAC nomenclature generation system for acyclic compounds. The program performs sequentially to construct a plausible IUPAC name for a given structural formula. The procedure follows faithfully the authentic rules; i.e. characteristic group search, determination of principal group, determination of principal chain, determination and naming of substituents, and construction of full name. The program has been tested for over one thousand structural formulas and obtained correct names for them with about 80% accuracy.

1. Introduction

In spite of the rapid progress of computer techniques for the manipulation of chemical structural formulas graphically, the primary importance of the systematic nomenclature has not decreased. If all compounds in the database are named uniquely, the most effective search key for a specific compound is its name.

At present, however, this is very difficult to realize because there is some historical confusion in the chemical-naming systems used by organic chemists. To overcome this confusion, the use of systematic nomenclatures such as IUPAC or CAS has been recommended. However, the naming rules for both these methods are very complicated and are difficult to learn even for chemists. Therefore, an automatic naming system is required in the field of database compilation. For CAS nomenclature, several papers concerning the development of an automatic naming system have been published [1,2], but with regard to IUPAC nomenclature, no notable papers have appeared.

The general idea of the present nomenclature system under development is shown in fig. 1. At present, the development of three modules, i.e. characteristic group search, chain nomenclature, and complete name construction modules, has been completed. Other modules, i.e. two ring naming modules and a stereo naming module, are not yet completed. Therefore, the present paper describes the details of the computer program for chain nomenclature.

As shown in fig. 2, there are seven alternative naming principles allowed in the IUPAC nomenclature system [3]. They are substitutive, radicofunctional, additive, subtractive, conjunctive, replacement, and assemble.

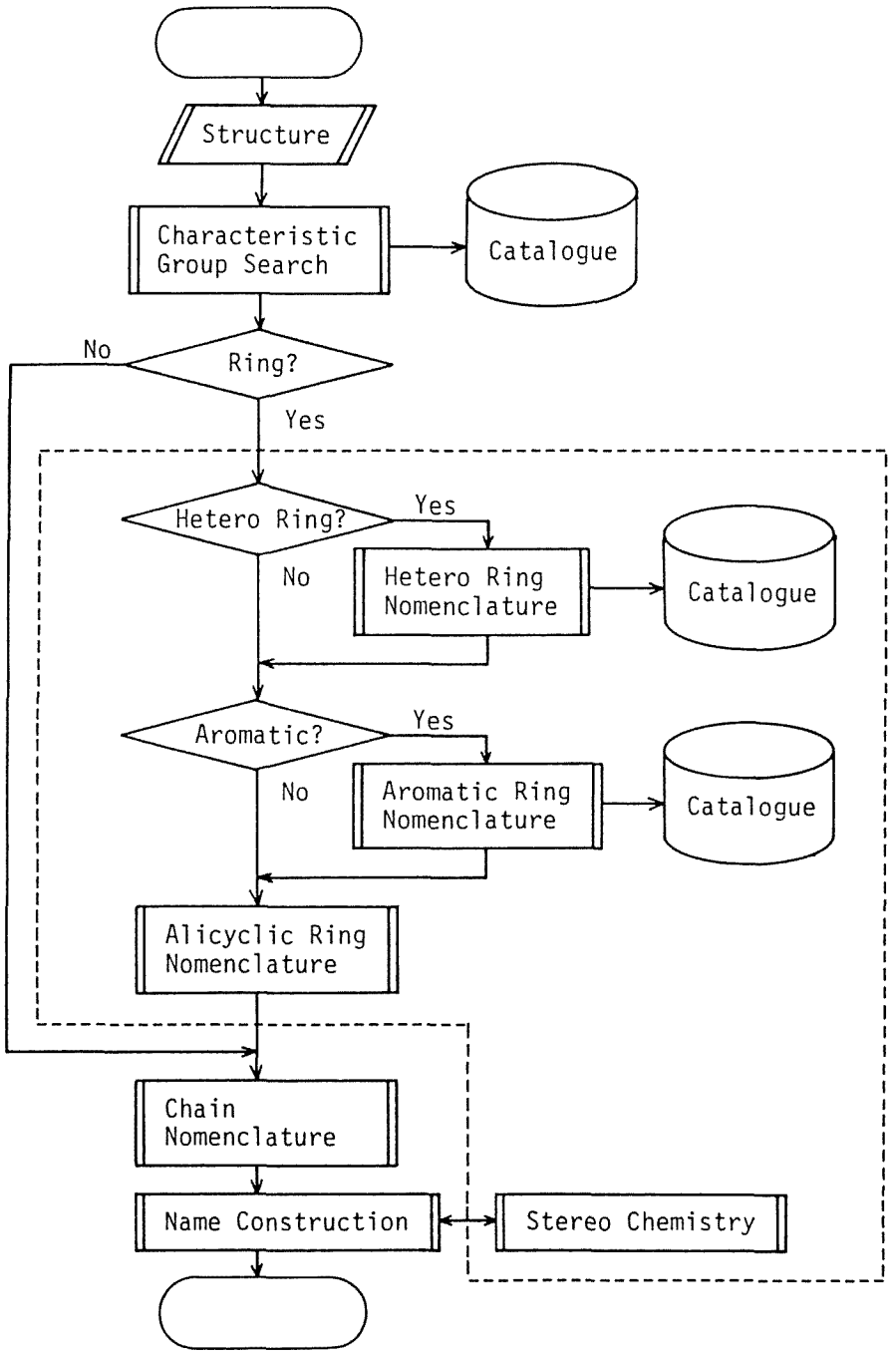


Fig. 1. General block diagram of the IUPAC nomenclature system.

1. Substitutive

$$\text{Structure} = [\text{Substituent}]_n + \text{Principal Chain} + [\text{Principal Group}]_m$$

$$\text{Name} = [\text{Prefix}]_n + \text{Principal Chain Name} + m[\text{Suffix}]$$

2. Radicofunctional 3. Additive 4. Subtructive

5. Conjunctive 6. Replacement 7. Assemble

Fig. 2. General concept of the IUPAC nomenclature system.

Since it is almost impossible to cover all the principles in a computerized system, the most general one, substitutive nomenclature, is selected as the target of automation. However, for some types of structure which could not be named by the substitutive principle, other principles are employed as an exception. Frequently appearing examples for such exceptions are names for alkyl amines and esters.

The radicofunctional principle is applied for the amines, and the ester is treated as a special case for the substitutive principle.

In the substitutive principle, a chemical structure is regarded as a combination of substituents, principal chain and principal groups, as shown in fig. 2. Then, the name consists of prefixes, principal chain name, and suffix. The names constructed by the present system are consistent with this structure.

General procedures for the formation of an IUPAC name from a given structural formula [3] are as follows:

(1) Search for all characteristic groups.

(2) Determination of the kind of characteristic groups for use as the principal group.

(3) Determination of the parent structure (principal chain).

(4) Determination and naming of the substituent(s).

(5) Naming of the parent structure including the principal group.

(6) Assembly of the partial names into a complete name.

The strategy taken in the present program system for naming various acyclic structures follows this procedure.

2. Characteristic group search and determination of the principal group

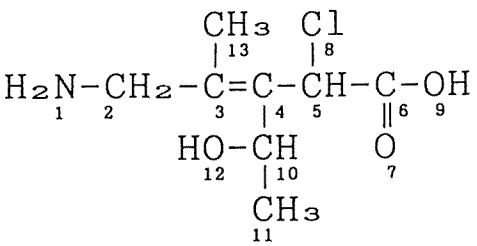
Examples of file records for characteristic groups are shown in fig. 3. As shown in the figure, ID number, number of atoms, seniority of the group, connection table, and some additional information are stored for each group. As for the seniority, a smaller numeral means a higher seniority.

```
1   Carboxylic acid                    ID Number, Group Name
3   1   1010                           Number of atoms, KSTAR, Seniority

8   1   0   1   2   1                  Connection Table
6   0   0   2   1   1   3   2
8   0   0   1   2   2

2   Peroxy acid
4   1   1020

6   0   0   2   2   1   4   2
8   0   0   2   1   1   3   1
8   1   0   1   2   1
8   0   0   1   1   2
    :
    :
    :
```

Fig. 3. Contents of the characteristic group catalogue.

Characteristic groups which are manageable in the present system are listed in table 1. This table shows the groups and their names as suffixes or prefixes. These two names are used for naming the given structural formula.

An example of input structure is shown in fig. 4. This structural formula is input as a form of connection table. At present, a more convenient graphical input system is also available. Using this structure as a working example, the naming procedures carried out in the present system will be explained in the following paragraphs.




```
13
 7   1   2   1
 6   2   1   1   3   1
 6   3   2   1   4   2  13   1
 6   3   3   2   5   1  10   1
 6   3   4   1   6   1   8   1
 6   3   5   1   7   2   9   1
 8   1   6   2
17   1   5   1
 8   1   6   1
 6   3   4   3  11   3  12   1
 6   1  10   1
 8   1  10   1
 6   1   3   1
```

Fig. 4. An example of input data.

Table 1

Characteristic groups and their names as suffixes or prefixes

ID	Atomic group	Seniority	Suffix	Prefix
1	–COOH	1010	–oic acid	carboxy
2	–COOOH	1020	peroxy..–oic acid	hydroperoxycarbonyl
3	–SO3H	1110	sulfonic acid	sulfo
4	–COOR	3010	R..oate	R-oxycarbonyl [oyloxy]
5	–COF	4010	–oyl fluoride	fluoroformyl
6	–COCl	4020	–oyl chloride	chloroformyl
7	–COBr	4030	–oyl bromide	bromoformyl
8	–COI	4040	–oyl iodide	iodoformyl
9	–CONH2	5010	–amide	carbamoyl [oylamino]
10	–CONHNH2	5110	–ohydrazide	carbazoyl [oylhydrazino]
11	–CHO	7010	–al	formyl
12	–CO–	8010	–one	oxo
13	–OH	9010	–ol	hydroxy
14	–OOH	20010	–	hydroperoxy
15	–NH2	10010	amine	amino
16	–NH–	10030	imine	imino
17	–O–	20110	–	R...oxy
18	–O–O–	20310	–	R...dioxy
19	–F	21010	–	fluoro
20	–Cl	21020	–	chloro
21	–Br	21030	–	bromo
22	–I	21040	–	iodo
23	–SO2H	1210	sulfinic acid	silfino
24	–SOH	1310	sulfenic acid	sulfeno
25		2010	reserved for future usage	
26	–C(=S)OH	1030	thioic acid	thiocarboxy
27	–C(=O)SH	1040	thioic acid	thiocarboxy
28	–C(=S)SH	1050	dithioic acid	dithiocarboxy
29	–S(=S)O2H	1120	thiosulfonic acid	thiosulfo
30	–S(=O)OSH	1130	thiosulfonic acid	thiosulfo
31	–S(=S)SOH	1140	dithiosulfonic acid	dithiosulfo
32	–S(=O)SSH	1150	dithiosulfonic acid	dithiosulfo
33	–S(=S)SSH	1160	trithiosulfonic acid	trithiosulfo
34	–S(=S)OH	1220	thiosulfinic acid	thiosulfino
35	–S(=O)SH	1230	thiosulfinic acid	thiosulfino
36	–S(=S(SH	1240	dithiosulfinic acid	dithiosulfino
37		2020	reserved for future usage	
⋮		⋮	⋮	
55		2280	reserved for future usage	

. . . continued

Table 1 (continued)

ID	Atomic group	Seniority	Suffix	Prefix
56	–C(=S)O–	3020	thioate	oxy(thiocarbonyl)
57	–C(=O)S–	3030	thioate	(R..thio)carbonyl
58	–SO2–O–	3110	sulfonate	oxysulfonyl [sulfonyloxy]
59	–SO–O–	3120	sulfinate	oxysulfinyl [sulfinyloxy]
60	–S–O–	3130	sulfenate	oxysulfenyl [sulfenyloxy]
61	–C(=S)–F	4050	thioyl fluoride	fluorothiocarbonyl
62	–C(=S)–Cl	4060	thioyl chloride	chlorothiocarbonyl
63	–C(=S)–Br	4070	thioyl bromide	bromothiocarbonyl
64	–C(=S)–I	4080	thioyl iodide	iodothiocarbonyl
65	–SO2–F	4110	sulfonyl fluoride	fluorosulfonyl
66	–SO2–Cl	4120	sulfonyl chloride	chlorosulfonyl
67	–SO2–Br	4130	sulfonyl bromide	bromosulfonyl
68	–SO2–I	4140	sulfonyl iodide	iodosulfonyl
69	–SO–F	4210	sulfinyl fluoride	fluorosulfinyl
70	–SO–Cl	4220	sulfinyl chloride	chlorosulfinyl
71	–SO–Br	4230	sulfinyl bromide	bromosulfinyl
72	–SO–I	4240	sulfinyl iodide	iodosulfinyl
73	–S–F	4310	sulfenyl fluoride	fluorosulfenyl
74	–S–Cl	4320	sulfenyl chloride	chlorosulfenyl
75	–S–Br	4330	sulfenyl bromide	bromosulfenyl
76	–S–I	4340	sulfenyl iodide	iodosulfenyl
77	–C(=S)NH2	5020	thiomide	thiocarbamoyl
78	–SO2–NH2	5030	sulfonamide	sulfamoyl
79	–SO–NH2	5040	sulfinamide	sulfinamoyl
80	–S–NH2	5050	sulfenamide	sulfenamoyl
81	–C(=S)NHNH2	5120	thiohydrazide	thiocarbazoyl
82	–SO2–NHNH2	5130	sulfonohydrazide	hydrazinosulfo
83	–SO–NHNH2	5140	sulfinohydrazide	hydrazinosulfino
84	–S–NHNH2	5150	sulfenohydrazide	hydrazinothio
85	–C(=NH)OH	5210	imidic acid	
86	–C(=NH)SH	5220	thiomidic acid	
87	–SO(=NH)OH	5230	sulfonimidic acid	
88	–S(=NH)OH	5240	sulfinimidic acid	
89	–C(=NNH2)OH	5310	hydrazonic acid	
90	–C(=NNH2)SH	5320	thiohydrazonic acid	
91	–SO(=NNH2)OH	5330	sulfonohydrazonic acid	
92	–S(=NNH2)OH	5340	sulfinohydrazonic acid	
93	–C(=O)NH–OH	5410	hydroxamic acid	
94	–C(=S)NH–OH	5420	thiohydroxamic acid	
95	–SO2–NH–OH	5430	sulfonhydroxamic acid	
96	–SO–NH–OH	5440	sulfinhydroxamic acid	
97	–C(=NOH)OH	5510	hydroximic acid	
98	–C(=NOH)SH	5520	thiohydroximic acid	
99	–SO(=NOH)OH	5530	sulfonohydroximic acid	

. . . continued

Table 1 (continued)

ID	Atomic group	Seniority	Suffix	Prefix
100	–S(=NOH)OH	5540	sulfinohydroximic acid	
101	–C(=NH)NH2	5610	amidine	amidino
102	–C(=NOH)NH2	5620	amide oxime	
103	–C(=NNH2)NH2	5630	amide hydrazone	
104	–C(=NNH2)NHNH2	5640	ohydrazide hydrazone	
105	–S(=NH)2OH	5710	sulfonodiimidic acid	
106	–S(=NH)(=NNH2)OH	5720	sulfonohydrazonimidic acid	
107	–S(=NNH2)2OH	5730	sulfonodihydrazonic acid	
108	–S(=NH)(=NNH2)OH	5740	sulfonohydroximimidic acid	
109	–S(=NOH)(=NNH2)OH	5750	sulfonohydrazonohydroximic acid	
110	–S(=NOH)2OH	5760	sulfonodihydroximic acid	
111	–CN	6010	nitrile	cyano
112	–OCN	6020		cyanate
113	–NCO	6030		isocyanato
114	–SCN	6040		thiocyanato
115	–NCS	6050		isothiocyanato
116	–C(=S)H	7020	thial	thioformyl
117	–C(=SO2)H	7030	thial dioxide	
118	–C(=SH	7040	thial oxide	
119	–C(=S)–	8020	thione	thioxo
120	–CO–	8110	one	oxo
121	=C=S	8120	thione	thioxo
122	–C(=SO2)–	8210	thione dioxide	
123	–C(=SO)–	8220	thione oxide	
124	–SH	9020	thiol	mercapto
125	–NHNH2	10020	hydrazine	hydrazino
126	–N=NH	10040	hydrazone	hydrazono
127	–S–	20120	–	thio
128	–S–S–	20320	–	dithio
129	–SO3	20210	–	sulfonyl
130	–SO2	20220	–	sulfinyl
131	–CO–COOH	20510	–	oxalo
132	H2N–C(=NH)NH2	5601	guanidine	guanidino
133	=N2	22010	–	diazo
134	–N3	22020	–	azido
135	–NO	22040	–	nitroso
136	–NO2	22040	–	nitro
137	–NH–O–	10100	hydroxylamine	–
138	–N2–	22010	–	azo
139	H2N–CO–NH2	10110	urea	ureido
140	HO–C(=NH)NH2	10120	isourea	isoureido

Each member of a characteristic group catalog is searched for the given structural formula. The search technique employed for this procedure is the graph-matching algorithm presented by Sussenguth [4]. A simple example is presented to illustrate the concept of this algorithm. Two graphs G and H, the former being one of the characteristic groups in the catalog and the latter the input structure to be named, are shown in fig. 5. The problem is to determine whether or not G is a subgraph of H and, if it is, to delineate the explicit correspondence between nodes.

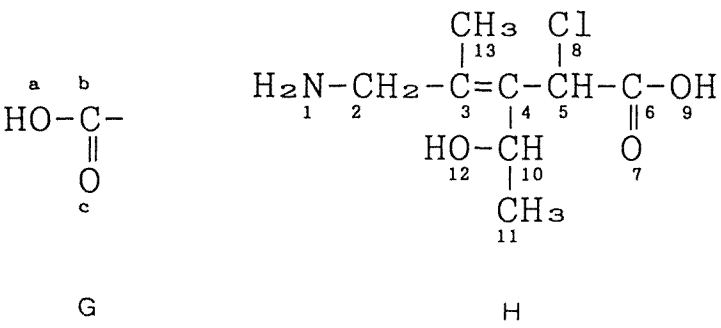


Fig. 5. An example for characteristic group search.

It is first examined whether those nodes of G which represent oxygen atoms should correspond to those nodes of H which represent oxygen atoms. That is, node a should correspond to either 7, 9 or 12 and no others, and also, node c should correspond to either node 7, 9 or 12. This is symbolically represented as $[a, c] = [7, 9, 12]$. Thus, considering all node values and the bonding information, the correspondences shown in lines 1–4 of fig. 6 are constructed.

Node			
C	[b]	[2,3,4,5,6,10,11,12]	line 1
O	[a,c]	[7,9,12]	line 2
Bond Order			
single	[a,b]	[1,2,3,4,5,6,8,9,10,11,12,13	line 3
double	[c]	[3,4,6,7]	line 4
Partition line 1 - 4			
	[a]	[9,12]	line 5
	[b]	[2,3,4,5,6,10,11,13]	line 6
	[c]	[7]	line 7
Connectivity			
line 7	[b]	[6]	line 8
line 8	[a,c]	[7,9]	line 9
Partition line 5 - 9			
	[a]	[9]	line 10
	[b]	[6]	line 11
	[c]	[7]	line 12

Fig. 6. Graph-matching procedure for searching a characteristic group in a given structural formula.

The sets on line 4 state that node *c* should correspond to either node 3, 4, 6 or 7. The correspondence concerning node *c* is also shown by the sets on line 2. To satisfy both requirements, node *c* should correspond to node 7.

Using this information, lines 5–7 are generated. Line 7 states that node *c* corresponds to node 7; it is now observed that node *b*, which is connected to node *c*, should correspond to node 6, which is connected to node 7. Using this information, the correspondence stated in line 9 is generated. Finally, exact correspondences in lines 10–12 are obtained, and graph *G* is determined to be a subgraph of graph *H*.

Atomic Group	Seniority	ID No.	Node Number												
			1	2	3	4	5	6	7	8	9	10	11	12	13
-COOH	1010	1	0	0	0	0	0	1	1	0	1	0	0	0	0
-CO-	8010	12	0	0	0	0	0	1	1	0	0	0	0	0	0
-OH	9010	13	0	0	0	0	0	0	0	0	1	0	0	0	0
-OH	9010	13	0	0	0	0	0	0	0	0	0	0	0	1	0
-NH2	10010	15	1	0	0	0	0	0	0	0	0	0	0	0	0
-Cl	21020	20	0	0	0	0	0	0	0	1	0	0	0	0	0

Fig. 7. Result of characteristic group search.

Figure 7 shows the result of a characteristic group search. The 6 groups presented in the figure were chosen from the catalog as the candidate characteristic groups.

There may be some overlappings among the remaining characteristic groups. For example, the second and third groups are, respectively, part of the first group. In such cases, the matching procedure described above is applied and those groups which are determined as subgraphs of others are deleted from the list.

However, in the case where the priority value of the smaller group is higher than that of the larger group, the larger group is deleted. For this example, both the second and the third groups are deleted, since they are subgraphs of the first one and have lower priorities.

Then, the group which has the highest seniority is selected as the principal group of the given structure. In this case, the first group, the carboxy group, has the highest priority and it is selected as the principal group and the remaining three groups are regarded as substituents.

For those structures which have no characteristic groups having suffix names in table 1, the parent names as hydrocarbons will be given.

3. Determination of principal chain

The next task to be done is the determination of the parent structure, which is called the principal chain for an acyclic structure. The following is a summary of rule C-13.1 of the IUPAC nomenclature:

(1) Maximum number of substituents corresponding to the principal group.

(2) Maximum number of double and triple bonds.

(3) Maximum length.

(4) Maximum number of double bonds.

(5) Lowest locants for the principal goups.

(6) Lowest locants for multiple bonds.

(7) Lowest locants for double bonds.

(8) Maximum number of substituents cited as prefixes.

(9) Lowest locants for all substituents in the principal chain cited as prefixes.

As described above, the provisions define what is the principal chain and the present algorithm follows this definition faithfully.

The three matrices shown in figs. 8 to 10 are made for the principal chain searching procedure. The first one is a distance matrix concerned only with carbon atoms in the given structure, as shown in fig. 8.

The second one is a multiple-bond matrix, which concerns the carbon–carbon multiple bond (fig. 9). The element M_{ij} of this matrix indicates the kinds and numbers of multiple bonds between nodes i and j as follows:

$$M_{ij} = (\text{no. of multiple bond}) \times 100 + (\text{no. of double bond}). \quad (1)$$

For example, the numeral 101 represents that there is one multiple bond and that it is a double bond on the chain designated by node 2 and node 4, and so on.

The last matrix is the principal group matrix. Each element of this matrix indicates the number of principal groups on a chain designated by two nodes, as shown in fig. 10. For example, there is one principal group on the chain designated by nodes 2 and 6, and so on. By means of these matrices, the selection of the principal chain is performed.

The procedure for searching the principal chain is illustrated in fig. 11. From the principal group matrix, all chains which have the maximum number of principal groups are chosen.

The maximum number is one for this case, and fifteen chains in the first column of the figure satisfy this condition. The first chain 2–6 and the fifth chain 6–2 are regarded as different for their locants on the nodes, and so on.

From these candidates, those which have the largest number on the third digit of the elements in the multiple-bond matrix are selected. In this case, six chains, 2–6, 3–6, 6–2, 3–6, 6–13 and 13–6, remain.

Then, by means of the distance matrix, the longest chains are searched for. As a result, the four chains 2–6, 6–2, 6–13 and 13–6 are selected because they all have the longest length, 4.

	1	2	3	4	5	6	7	8	9	0	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	2	3	4	0	0	0	3	4	0	2
3	0	1	0	1	2	3	0	0	0	2	3	0	1
4	0	2	1	0	1	2	0	0	0	1	2	0	2
5	0	3	2	1	0	1	0	0	0	2	3	0	3
6	0	4	3	2	1	0	0	0	0	3	4	0	4
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	3	2	1	2	3	0	0	0	0	1	0	3
11	0	4	3	2	3	4	0	0	0	1	0	0	4
12	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	2	1	2	3	4	0	0	0	3	4	0	0

Fig. 8. Distance matrix for carbons.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	101	101	101	0	0	0	101	101	0	0
3	0	0	0	101	101	101	0	0	0	101	101	0	0
4	0	101	101	0	0	0	0	0	0	0	0	0	101
5	0	101	101	0	0	0	0	0	0	0	0	0	101
6	0	101	101	0	0	0	0	0	0	0	0	0	101
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	101	101	0	0	0	0	0	0	0	0	0	101
11	0	101	101	0	0	0	0	0	0	0	0	0	101
12	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	101	101	101	0	0	0	101	101	0	0

(no. of multiple bond) x 100 + (no. of double bond)

Fig. 9. Multiple-bond matrix.

From the remaining four chains, those which have the maximum number of double bonds are searched by means of the last digit of the elements in the multiple-bond matrix.

The procedure described above corresponds to the first four selection rules listed previously. For this example, the four chains are equally possible candidates for the main chain. Then, the selection procedure is continued further by applying the remaining selection rules.

	1	2	3	4	5	6	7	8	9	0	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0	0	0	0
6	0	1	1	1	1	1	0	0	0	1	1	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	1	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	1	0	0	0	0	0	0	0

Fig. 10. Principal group matrix.

P.G. Matrix	M.B. Matrix	D. Matrix	M.B. Matrix
2 - 6	101	4	101
3 - 6	101	3	
4 - 6			
5 - 6			
6 - 2	101	4	101
6 - 3	101	3	
6 - 4			
6 - 5			
6 - 6			
6 -10			
6 -11			
6 -13	101	4	101
10 - 6			
11 - 6			
13 - 6	101	4	101

Fig. 11. Search procedure for principal chain – 1.

The further selection procedure is illustrated in fig. 12. Those four chains shown in the first section of the figure are the candidates selected in the previous step.

The first selection is made on the basis of the locant of the principal group. The two chains shown in the second section of fig. 12 remain because both have locant 1 for the principal group.

Candidates of Principal Chain

2-3-4-5-6
6-5-4-3-2
6-5-4-3-13
13-3-4-5-6

Selection by Seniority of Locant of P.G.

6-5-4-3-2 locant 1

6-5-4-3-13 locant 1

Selection by Seniority of locant of M.B.

6-5-4-3-2
0 0 101 0 locant 3

6-5-4-3-13
0 0 101 0 locant 3

Selection by Number of Substituents

6-5-4-3-2
0 1 1 1 1 total 4

6-5-4-3-13
0 1 1 1 0 total 3

Fig. 12. Search procedure for principal chain – 2.

The next selection is made on the basis of the locant of the multiple bond. In this case, there is only one double bond for each chain, and both have locant 3 and they still tie.

Then the next selection is made on the basis of the total number of substituents on the chain. Among these two, the first chain has four substituents and is superior to the second one, as shown in the bottom section of the figure.

Then, the chain 6–5–4–3–2 is selected as the principal chain. If plural candidates still remain, they are treated as equally possible candidates.

4. Determination and naming of substituents

The next step of the naming procedure is the determination and naming of the side chain with/without subsidiary substituents. In the present algorithm, those

side chains which do not strictly coincide with any members of the characteristic group catalog are regarded as the object of this procedure. Therefore, a major part of this procedure consists of the naming of carbon chains, i.e. alkyl radicals.

The construction of the name of monovalent carbon chain radicals can be represented as shown in fig. 13. Concerning the polyvalent radicals, only those which have all valence bonds on the same carbon atoms can be named with the present program.

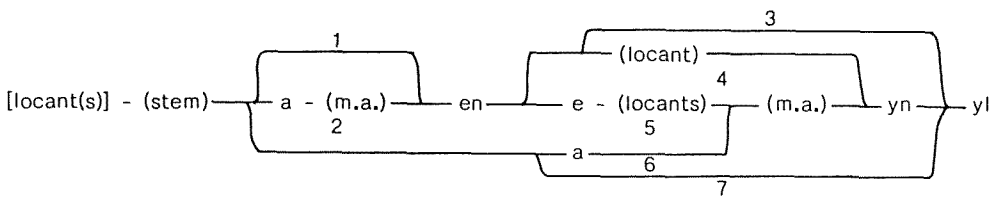


Fig. 13. Schematic diagram of monovalent alkyl radical nomenclature.

The term "locant(s)" in the figure means the numeral(s) which indicates the position(s) of the unsaturated bond(s), "stems" means IUPAC naming stems for straight carbon chains, and "m.a." means multiplying affixes such as di, tri, tetra, penta, and so on.

The "stem" for a specific chain is determined by referring to the previously defined table. The table contains the length of the carbon chains and corresponding terms, for example, "meth" for length 1, "eth" for length 2, "prop" for length 3, "but" for length 4, "pent" for length 5, etc. It can be enlarged virtually to any length, but at present 50 is the limit of the chain length.

For the present example, the side chain consists of carbons no. 5 and 6, and oxygen no. 12 is the case in point.

A similar procedure for searching the parent chain as previously described is employed for searching the parent substituent chain, too. The length of the main chain is 2, then the stem name "eth" is chosen from the stem table and the suffix "yl" is attached to the stem to make the side chain name "ethyl". In this case, the naming route 7 in fig. 13 was taken. Then, the name of this side chain is determined as "1-hydroxyethyl", as shown in fig. 14. This routine can be applied recursively for more complicated side chains.

An additional example is shown in fig. 15. For this structure, the parent substituent chain is determined as 1–2–3–4–5–6–7–8–9–10 according to the algorithm described previously. Then, the subsidiary substituent chain 11–12–13–14–15 is named first:

The length of the chain is 5, therefore the stem name is "pent".

The number and kind of multiple bond is one double bond.

The locant of the double bond is "3".

Substituent to be named as prefix:	10, 11, 12
Parent Substituent Chain:	10-11
Subsidiary prefix:	12
Name of Subsidiary Prefix:	1-hydroxy
Name of Parent Substituent Chain:	ethyl
Complete Substituent Name:	1-hydroxyethyl

Fig. 14. Naming procedure for side chain.

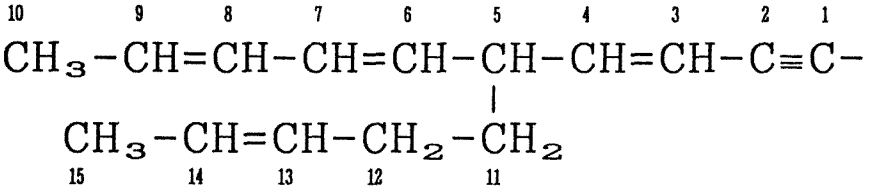


Fig. 15. An example for alkyl nomenclature.

Thus, the naming route 1–3 in fig. 13 is taken as 3-pent-en-yl and the name 3-pentenyl is obtained.

Secondly, the parent substituent chain is named as follows:

The length of the chain is 10, therefore the stem name is "dec".

The numbers and kinds of multiple bonds are three double bonds and one triple bond, respectively.

The locants of the double bonds are 3, 6 and 8, and the locant of the triple bond is 1.

Thus, the naming route 2–4 in fig. 13 is taken as 3, 6, 8-dec-a-tri-en-1-yn-yl and the name 3,6,8-decatrien-1-ynyl is constructed.

Finally, the complete substituent name 5-(3-pentenyl)-3,6,8-decatrien-1-ynyl is constructed. Hyphenation and insertion of parentheses are also executed properly. Details of the procedures will be presented in the near future.

There are many trivial names which are preferred or allowed in the IUPAC nomenclature system; the representative examples are small alkyl radical names cited in rules A-2.25 and A-3.5, and fatty acid names. For the former cases, automatic conversion is performed after completing the naming procedure. For example, if a character string "1-methylethyl" is found in the given name, then it is converted into "isopropyl", and so on.

5. Naming of the parent structure

The next step is the naming of the parent structure including the principal group, as shown in fig. 16. The principal chain has been determined as 6–5–4–3–2 in the previous step. The location, the number and the kind of multiple bonds

Principal Chain:	6-5-4-3-2
Multiple Bonds:	1 double bond at position 3
Name for Principal Chain:	3-pentene
Suffix for Principal Group:	-oic acid
Complete Name for Parent chain:	3-pentenoic acid

Fig. 16. Naming procedure for principal chain.

have also been determined. A similar procedure for naming carbon side chains is employed here, the only difference for the principal chain being the suffix. So, the name of the principal chain is obtained as "3-penten(e)".

The suffix of the principal group had already been obtained from the catalog as "oic acid". Therefore, the complete name of the parent chain is constructed by deleting the ending "e" from the principal chain name and connecting this suffix as "3-pentenoic acid".

6. Construction of the complete name

Now, all parts for constructing the complete name are at hand. These are the parent name, and all substituent names and their locants. Then, the substituent names with locants are arranged in alphabetical order in front of the principal chain name and give the complete name for the structure, as shown in fig. 17.

Parent Chain		Name
C-C=C-C-COOH		3-pentenoic acid
Substituents	Position	
-NH2	5	amino
-CH(OH)-CH3	3	(1-hydroxyethyl)
-CH3	4	methyl
-Cl	2	chloro

Complete Name

5-Amino-2-chloro-3-(1-hydroxyethyl)-4-methyl-3-pentenoic acid

Fig. 17. Construction of complete name.

In actual fact, several procedures for inserting appropriate parentheses and hyphens have been executed for constructing the complete name of the given structural formula.

7. Results and discussion

Some examples for which the present program gives different names to those which appeared in the Dictionary of Organic Compounds (DOC) [5] are shown in fig. 18. For the first two examples, both names are reasonable for the structures. However, for the third example the name appearing in the dictionary is incorrect. In this case, the principal chain should contain a double bond. Thus, the correct parent chain should be butenoic acid.

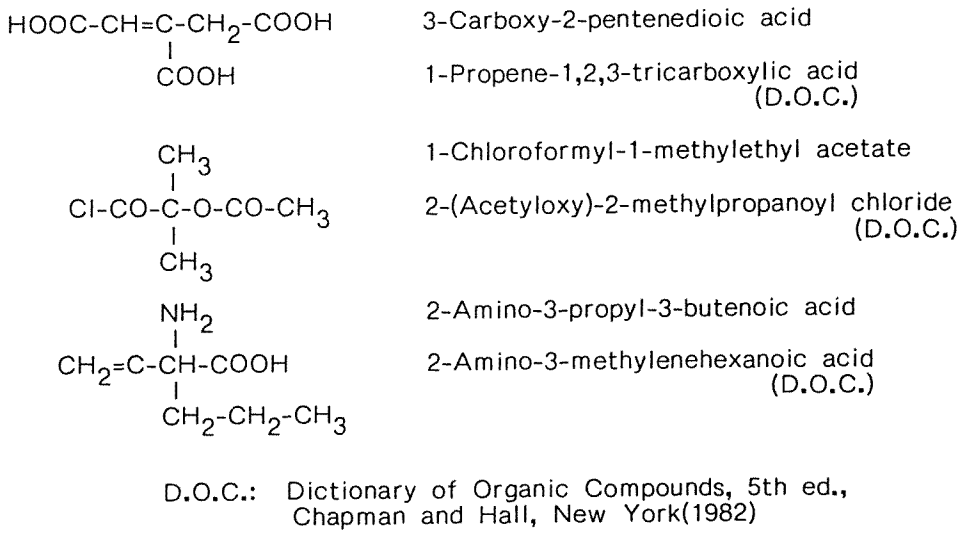


Fig. 18. Examples of the results obtained by the naming system.

Additionally, the present system has been tested for over one thousand acyclic structures appearing in the DOC, and correct names were obtained for approximately 80% of them. Most of the compounds for which the present system could not give correct names contain nitrogens. It can be said that IUPAC nomenclature rules for nitrogen containing structures are not systematic, but too empirical.

At present, IUPAC naming for acyclic structures only can be done with the present system. However, the development of naming algorithms for cyclic structures is almost completed and will be presented in the near future.

References

[1] J. Mockus, A.C. Isenberg and G.G. Vander Stouw, J. Chem. Inf. Comput. Sci. 21(1981)183.
[2] D.E. Meyer and S.R. Gould, Amer. Lab. (Nov. 1988)92
[3] J. Rigaudy and S.P. Kleney, *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F and H, 1979 Edition* (Pergamon Press, Oxford, 1979).
[4] E.H. Sussenguth, Jr., J. Chem. Doc. 5(1965)36.
[5] J. Buckingham (ed.), *Dictionary of Organic Compounds*, 5th Ed. (Chapman and Hall, New York, 1982).