

4G-7

化合物名称解析サブシステム

・・朴 光実 ・・伊藤 照明 ・大保 信夫 ・鈴木 功 ・藤原 譲

・・筑波大学理工学研究科

・筑波大学電子情報工学系

1. はじめに

膨大なデータの集積がなされている化合物情報の中心をなす化学構造の正確な入力、表現は専門の化学研究者にとっても困難であるので、それを支援するシステムとして化合物名称解析サブシステムを目標としている。

これまで CAS (米国)、日本科学技術情報センター (JICST)、Beilstein (独) などでのこの種のシステムが開発され、名称から構造情報への変換を行っている。それらはそれぞれシステム固有の命名法に基づいている。本システムでは国際的に使用されている各種の命名法に対処できる実用的な機能を持つシステムの開発を目指す。特に国際的標準命名法である IUPAC の名称と構造情報の双方向変換を行う機能を実現し、多様な同意語の取り扱いや、研究開発に重要な総称表現の処理ができるシステムの開発を目標としている。これにより命名規則の整合性、完全性の検証を通じて IUPAC 命名法の改善も可能となる。

2. IUPAC 命名法

本研究は、IUPACS PROJECT(The Integrated Utility Package for Advanced Chemical Research Systems) — 標準化合物情報の統合システム・プロジェクトの化合物名称解析サブシステムである。このプロジェクトは化学の研究に必要な化合物の名称、構造、(トポロジカル、二次元および三次元)を蓄積、検索、変換、表示する機能を持つ高度な研究開発用統合化合物情報データベースシステムを構築するというものである。化合物名に対して、IUPAC 命名法に基づいた命名規則に従って、化合物の個別名称、総称名称を解析したうえで構造式に変換し、また、同じ解析辞書を利用して構造式から名称への変換も可能である。

IUPAC 命名法というのは国際純正及び応用化学連合 (International Union of Pure and Applied Chemistry) による標準命名法である。炭化水素及び基本複素環を母体とし、母体化合物名と特性基名との組合せによる命名法を中心とする各種の命名法から構成されており、そのルールの体系は極めて複雑である。これによって命名された化合

物名は、文字列から母体名や、置換基名などの辞書を使って矛盾が生じないように部分名称を取り出し、それぞれについて化学式や結合関係について解析し、対応する構造式に変換する。

3. 有機化合物 IUPAC 組織名称構文解析:

IUPAC 命名法では、構造のわかった化合物に対して、なるべく簡単でしかも一意的に化合物の構造を表しうるような名称を定めることを目的としている。この目的に沿って、特別に作られた名称を組織名と呼ぶ。組織名の解析のため、その構成を次のように整理した。

構造:: 置換基 + 母体構造 + 置換基

名称:: {「接辞」+「接頭語」}n + 「接辞」+ 母体名 + 「接辞」+「接尾語」

母体名:: 「語頭」+「接辞」+語幹+「接辞」+語尾

語幹: 語幹は骨格炭素の数を表す。

(語幹辞書)

語尾: 骨格炭素の間の結合種類を表すものである。(語尾辞書)

語頭: 母体の型を表すものである。

(語頭辞書)

接辞: 接する部分語の位置、数や状態を示すのにつかわれる位置番号や倍数接語などである。

位置番号: アラビア数字、ローマ字、ギリシャ文字である。

(接辞辞書)

倍数接語: 同種の置換基を示す英語名などである。(接辞辞書)

接尾語: 母体名のあとにおく置換基である。1つの母体名に対して1つの接尾語しかない。接尾語置換基間には順位がある。(置換基辞書)

接頭語: 母体名の前におく置換基である。接頭語の数は制限しない。接頭語置換基間には基名の頭

文字のABC順位に並べる。(置換基辞書)

置換基: 炭化水素あるいは基本複素環系の水素原子置換している原子または原子団を置換基と総称する。

ここで「」内は名称中に存在しないこともあることを示す。 x_n は x を $n \geq 0$ 回の繰り返す意味である。

4. 組織名称解析アルゴリズム:

① 名称を左から下の文法(1)によって、1キャラクタずつスキャン(scan)する。文法(1)を満足する形態素を図・1のようにリストにいれる。[と(のある形態素に対しては、①を繰り返す。

文法(1):

```
< subname > ::= 「 < D > { < E > | < F > } 「 - 」
< E > ::= < L > 「 < C > | < F > 」
      { < L > 「 - 」 < L > } *
< F > ::= [ < N > { { { < N > } |
      { " , < N > " } } * | < C > } ]
< C > ::= ( { < N > | < L > | , | - } * )
< D > ::= < N > 「 ' 」 { , < N > 「 ' 」 } * -
< L > ::= a | b | c | ... | z | . | " | '
< N > ::= 0 | 1 | 2 | ... | 9 | P | H | A | B
```

ここで < > 内は構文要素 (syntactic element) である。「」は同上の意味である。| は「または」という意味である。 x^* と x^+ はそれぞれ x を $n \geq 0$ 、 $n > 0$ 回の繰り返す意味である。「_」は名称の中の空のかわりに使うものである。{ } 内は名称中に必ず存在するという意味である。

② 辞書マッチング

まずリストから辞書によって母体名を捜す。母体構造が決ってから母体の各ノードに結合されているものを決める。(図・2)

③ 結合表の組立 (図・3)

5. おわりに

本研究は、IUPAC の名称と構造情報の双方向変換を行うことや総称表現を取り扱うなど、従来のシステムにはなかった機能が盛り込んで、国際的に使用されている各種の命名法に対処できる実用的な機能を持つシステムを目指して開発中である。

参考文献

- 1) 藤原 他: 「ネットワーク共用による化合物情報等の利用高度化に関する研究における、化合物辞書システム・立体化学的物質登録」、「立体化学的物質登録データの蓄積と機械処理向き体系名規則の作成」 情報科学技術研究会発表論文集 (P.P.49~60, P.P.75~84, 1984)

名称: (文字列)

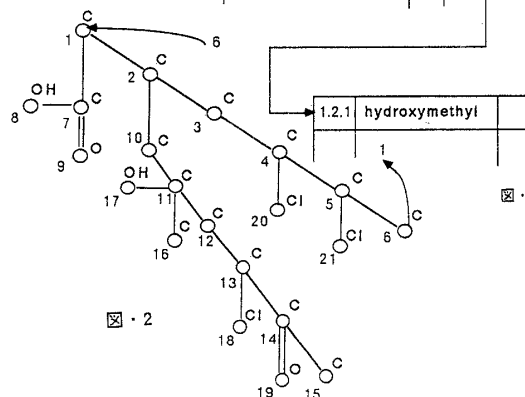
4,5-dichloro-2-[4-chloro-2-(hydroxymethyl)-5-oxohexyl]-1-cyclohexanecarboxylic_acid

形態素: (文法 1)

0	4,5-dichloro-
1	2-[4-chloro-2-(hydroxymethyl)-5-oxohexyl]-
2	1-cyclohexanecarboxylic_acid

辞書 matching 結果:

1.1	4-chloro-
1.2	2-(hydroxymethyl)
1.3	5-oxohexyl

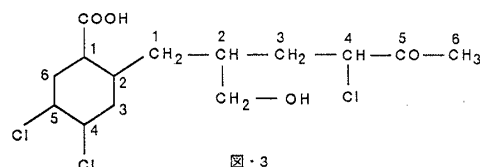


図・1

結合表 (Connection Table):

番号	原子	結合	(種類 相手)	座標(x y)
1	C	2	12 17 0 0	0 0
2	C	2	13 110 0 0.66	-0.33
3	C	1	14 0 0 0.66	-1.33
4	C	2	15 120 0 0	-1.66
5	C	2	16 121 0 -0.66	-1.33
6	C	1	11 0 0 -0.66	-0.33
7	C	2	18 29 0 0	1
8	O	0	0 0 0 0	2
9	O	0	0 0 0 0	-1
10	C	1	111 0 0 1.33	0
11	C	2	112 116 0 2.33	0
12	C	1	113 0 0 3.33	0
13	C	2	114 118 0 4.33	0
14	C	2	115 219 0 5.33	0
15	C	0	0 0 0 6.33	0
16	C	1	117 0 0 2.33	-1
17	O	0	0 0 0 2.33	-2
18	Cl	0	0 0 0 4.33	-1
19	O	0	0 0 0 5.33	-1
20	Cl	0	0 0 0 0	-2.33
21	Cl	0	0 0 0 -1.33	-1.66

構造:



図・3

2) Dittmar, Mockus, and Couvreur: "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams" J.Chem.Inf.Comp.S., Vol.17, No.3, 1977

3) 内野浩、荒木啓介: 「有機化合物の体系的名称から結合表、GREMASコードおよびディスクリプタへの自動変換」 情報科学技術研究会発表論文集 (P.P.101~121, 1977)