

化学構造情報の処理

内田 尚子 (日本科学技術情報センター)

1. はじめに

化学は物質の構造, 性質, 化学反応等を扱う学問であるという性格上, 化学文献の中に含まれる情報は, 化合物を種々の側面から取り扱ったものが多い。新しく抽出または合成される化合物も毎年30万を下らないといわれ, 化合物ファイルとして世界最大の CAS Registry File に登録された化合物数は, 既に350万前後に達している。

化合物の中でも有機化合物は種類が多く, 全構造から細々の構成元素に至るまで, その構造の部分部分がすべて興味の対象となる可能性をもっている。その蓄積検索システムは, できるだけ蓄積容量を小さく保ちながら, 多様な検索要求を満足させるような豊富な情報を内蔵していなければならない。ここでは, 化合物情報の全体像を紹介してから, JECSTにおける化学構造情報処理について述べる。

2. 化合物情報の構成

2.1 化合物の数と種類

収録化合物数では最大規模の CAS Registry File を例にとり, ファイルのサイズ, 平均年間増加量を第1表に, 収録化合物の種類とその数を第2表に示す。

第1表. CAS Registry File のサイズ⁽¹⁾.
(1974年5月のデータ)

	化合物数	サイズ(バイト)	平均増加バイト/年
構造ファイル	2.8×10^6	870×10^6	114×10^6
名称ファイル	3.1×10^6	830×10^6	80×10^6

→ 30万化合物/年12相当する。

(1) Hufferberger, M.A., Wigington, R.L.: Chemical Abstracts Service Approach to Management of Large Data Bases, J Chem Inform Comp Sci 15(1), 43-47 '75

第2表 CAS Registry File 中の化合物の種類⁽²⁾
(1974年1月当時0270万化合物についての調査)

化合物種類	数
構造判明のもの	2,600,000
部分的に構造が不明確なもの	41,000
合金	32,000
ポリマー	84,000
混合物	7,000
配位化合物	178,000
イオン	
元素	
フリーラジカル	

2.2 化合物情報を表わすデータエレメント

化学関係の文献ファイルは、対象文献数の多い大規模なCAC, CASIAなどや、少数の文献から詳細に化合物情報を蓄積しているICRS, RINGDOCなど、数多く存在する。また化合物ファイルとしては前出のRegistry File やその部分に当るCHEMLINE ファイルがある。

これらのファイルの中では、化合物情報は単なる一般名から構造の大小様々な部分を組立てて作り上げる形式で表示し、種々のレベルの採集ができるような工夫がなされているもので、いろいろな表記法で蓄積されている。

第3表は、化合物情報が、各種のファイルの中で、どのようなデータエレメントとして含まれているかを示したものである。

2.3 構造特性の分布

化合物の検索手順を考える場合には、ファイル中の化合物の構造特性の分布がわかっていると便利である。第4表と第5表は1974年1月時点におけるCAS Registry File 中の化合物についての元素と環構造特性の分布を示している⁽²⁾。

(2) CAS: Substructure Searching of Computer-readable CAS 9CI Chemical Nomenclature Files. Aug. 1974.

表 3 化合物情報のデータエレメント

データエレメント名	内容 (例)
一般名	何々の化合物の一般的な呼称または商品名 (Bacitracin F)
半組織名	立体母核 + 組織名称フラグメントで構成される名称 (Gammaceran-3-ol)
組織名	
母核名	その化合物の官能基群のうち、優先性の最も高いものを示す名称と、その主官能基が結合している環式ないし非環式骨格の名称。配位化合物の中に元素名のこともある (Benzeneethanamine)
同形異義識別子	母核名が同形異義語群の一つである場合に、そのちがいを表示する語句または分子式など (Serpentine --- alkaloid " --- mineral)
Line Formula	化学物質をさらに詳しく記述するの に用いられる 1 行表示の分子式 (Copper chloride --- $CuCl_2$)
置換基	母核名に直接または間接に結合している従属部分構造を示す名称接頭辞、または有機配位子 (4-chloro- α -methyl-)
名称修飾辞	構造を完全に定義するため追加が必要となる特定の化学誘導体情報を示す (hydrochloride)
立体化学記述子	他のデータ要素で記述されていないような立体化学情報を示す (+) -
分子式	化合物全体の元素組成を示す
原子結合行列	
結合	原子間の結合 (グラフと辺) の種類
グラフ	構造中の原子間の結合関係を表わす行列
節	グラフの頂点の原子の元素名
環数	化合物に含まれる環の数
環のサイズ	環系を構成する各環の員数
環の元素構成	" " " 元素 (H 以外) 組成 (C3N2)
環成分元素の線状表記	2 個以上の非炭素原子をもつ環の構成元素を結合順に表記したもの (NCNC2, N2C3)

表3 (つづき)

データエレメント	内 容
WLN	末端グループを表わす記号の優先性の高い方から規則的にフラグメント記号を表示した化合物表示 ($\text{H}_2\text{N}-\text{CH}_2\text{CH}_2-\text{C}(=\text{O})-\text{OH} \rightarrow \text{Z2VQ}$)
GREMASコード	構造フラグメントを示す3けたのコードと、フラグメント間の結合を示すコードの組合せで全構造を表示する体系
その他	

表4 Registry File 中の化合物における元素の分布⁽²⁾

元 素 名	当該元素を含む化合物の割合 (%)
H	97.36
C	96.49
O	81.24
N	64.25
S	20.90
Cl	19.98
P	6.22
F	6.10
Br	5.39
Si	2.43
I	2.38
Na	1.51
Fe	1.48
Co	1.40
B	1.32
Ni	1.26
Cu	1.20

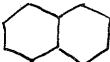
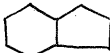
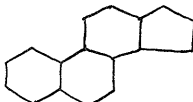
(注) ここには1%以上の元素を示した。

第5表 環構造の命題⁽²⁾

表5-1 非環式および環式化合物の分布

非環式	400,000
環式	2,300,000
環系 1 億	890,000
" 2 "	600,000
" 3 "	230,000
" 4 "	75,000
" 5 ~ 25 億	53,000

表 5-2 環型のりん度分布 (りん度順上位5種)

環型	度数
	2,362,000
	369,000
	216,000
	203,000
	67,400

(注) 環型の総数 : 14,450

環素の総数 : 101,900

環系の出現頻度計: 3,861,000

出現次數上達200種

の環型の出現頻度計: 3,720,000 (環系の出現率 96%)

環型と環系との関係:

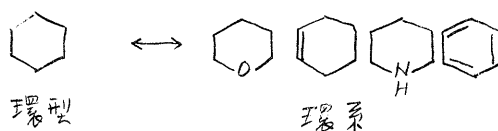
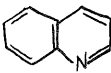
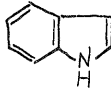
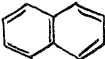
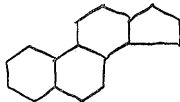
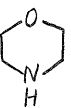
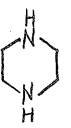


表5-3 環系 α 分布 (上位20種)

環系	出現数	環系	出現数
	1,861,000		25,100
	95,000		23,100
	80,000		22,600
	61,500		22,300
	57,000		18,500
	45,000		17,900
	40,300		17,100
	37,800		16,900
	28,000		
	27,800		
	27,100		
	25,500		

3. JICSTにおける化学構造情報処理

3.1 前提条件

- (a) JICST オンライン検索システム(JOIS)が公衆回線を通じて利用可能になると共に端末の数の増加率が高まることが予測される
- (b) 利用者自らも探索者である
- (c) 不特定多数の利用者を対象にするが、その多くは化学者自身である
- (d) 化学者が日常使いたれた化合物情報についての言語体系は構造式と名称である
- (e) 検索効率の最も高い検索結果は原子結合行列を対象としたトポロジカル探索によって得られるが、種々のスクリーンの組合せによる探索でほぼ満足できる結果を、はるかに迅速に得られることが実証されている。名称による探索はスクリーン探索の1つであり、トポロジカル探索の結果を参照(100%)とした平均再現率は97.9%、平均適合率は64.5%だったというデータがある

3.2 対象ファイル

文献ファイル中でフラグメントコードに相当する化合物ディスプレイによる化合物検索が可能なJICSTファイルが現在稼働中であり、化合物情報専用ファイルとして、Registry Nomenclature Fileの部分ファイルに相当するCHEMLINEファイルの導入をすすめている。CHEMLINEはNLMのTOXLINEファイル(毒物情報を含む文献ファイル)に含まれる化合物についての辞書ファイルであるが、今後化合物を抽出する対象文献ファイルの種類が増えることが予定されている。このファイルのユニットレコードの内容例を次の表に示す。

3.3 検索システムと検索用データエレメント

名称による部分構造探索は、既成の文献検索システムで行えるという利点がある。CHEMLINEファイルに関して、JICSTの現行の検索システムを利用する場合の1次タプルは、名称探索における1次スクリーンとして機能するが、本来この1次スクリーン探索で値内の多くが満足すべき結果を得ることが望ましい。最も重要なスクリーンである名称フラグメントを作るための名称データのフラグメント化については、名称データだけでなく、CACファイルのキーワードとタイトルワードも対象としたフラグメント化システムを現在作成中である。その他に、各種のデータエレメントのどれから、どのような形式のスクリーンを作り出するか、それに関連したマスターファイル(2次タプルおよび出力用データエレメントを含む)の構成については、今後検討をすすめる。

第6表 CHEMLINE 7 p 3 IL92277 TL2 - F^{1/34}



RN - 288-32-4
 MF - C3-H4-N2
 NI - IMIDAZOLE (8CI)
 NI - 1H-IMIDAZOLE (9CI)
 SY - 1,3-DIAZA-2,4-CYCLOPENTADIENE
 SY - 1,3-DIAZOLE
 SY - FORMAMIDINE, N,N'-VINYLENE-
 SY - GLYOXALIN
 SY - GLYOXALINE
 SY - IMIDAZOL
 SY - IMUTEX
 SY - MIAZOLE
 SY - METHANIMIDAMIDE, N,N'-1,2-ETHENEDIYL-
 WL - T5M CNJ
 NR - 1
 RS - 5
 RE - C3N2
 CL - NCNC2
 LO - TOXLINE