情報学基礎 1－5
（1986・7・24）

「化合物構造のデータベース化」

荒木啓介，梶正憲，木村美実子，前田知子，日夏健一，
日本科学技術情報センター（JICST），　清水英昭

既に600万以上もあると言われている化合物は、一般に複雑な構造を持ち、ま た微妙な構造上のちがいが、生物活性や物性に大きく影響することが多い。

これら化合物は、俗称である慣用名で呼ばれることもある外、正式には、一種の人工言語である体系名規則によって命名され、記述される。JICSTでは、科学技術庁、振興調整費プロジェクトの依託を受けて、この体系名から化合物構造の電子計算機表現形式の一つである分子構造の原子結合表を立体化学のレベルで自動的に組立てるプログラムと、その内部辞書である解析辞書を開発した。

化合物名データ1万についてテストを行い、ほぼ当初の設計どおりの結果を得ている。

CHEMICAL COMPOUND DATA BASE

ARAKI Keisuke, KAJI Masanori, KIMURA Mimiko, MAEDA Chikako, HINATSU Kenichi, SHIMIZU Hideaki

The Japan Information Center of Science and Technology　2-5-2 Nagatacho Chiyodaku Tokyo Japan

Chemical compounds, reported to amount to 6 million, have in general complicated structure and it's slight differences may affect greatly their biological or physico-chemical properties.

These compounds are either called by trivial names or officialy named according to nomenclature rules which are thought to be a kind of artificial languages. JICST, entrusted by Science and Technology Agency of Japan, developed computer programes and inner dictionary to automatically generate stereo-specific s atom connection table from almost all kinds of nomenclatures. The test for 10,000 compounds showed satisfactory results.

1、化合物体系名の要素の分析と、解析辞書項目

1.1 主構成要素

(1) 骨格

(a) 当初から辞書要素として，持たなければならない骨格

① 一連の鎖状炭化水素と鎖状ヘテロ化合物
② 一連の単環
③ 縮合環系
④ 単位となる糖，アミノ酸，ステロイド，テルペンなどの天然物の骨格
⑤ 一連の骨格と官能基を合せ持った名称（酢酸，メタノールなど）
⑥ その他，アルゴリズムによって，組み立て不可能か困難な単位

(b) アルゴリズムによって，組立て可能とする骨格の範囲

① 架橋化合物
② 異原子による原子置換を受けた骨格
③ スピロ化合物
④ 指示水素により規定された骨格
⑤ 脱水素あるいは水素化を受けた骨格
⑥ ホモ，ノル，セコ，シクロなどによる環骨格の変形

これらは，上記(a)の骨格をもとに，各種付加的操作をほどこす。

(2) 置換基

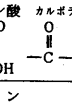
置換基の種類と分類

項番	分類コード	他原子(団)の結合を受けるか否か	メチル，メトキシへの結合	手の価数	手を有するノード数	例
1	S 10	受けない	可	1価以上	1	オキソ =O　ニトロ —NO$_2$　クロロチオ Cl—S—　メチルチオ CH$_3$—S—
2	S 20	受ける	可	2価以上	1以上	カルボニル —C(=O)—　チオ —S—　オキシチオ —O—S—
3	S 21	受ける	可	2価以上	1，2	イミノ —NH—　メチレン —CH$_2$—　マロニル —C(=O)—CH$_2$—C(=O)—　メタントリイル —CH<
4	S 22	受ける	可	1価	1	アミノ —NH$_2$　クロロシリル Cl—SiH$_2$—
5	S 23	アルキル基の置換は受けない。その他の原子(団)の置換は受ける。	可	1価	1	メトキシ —O—CH$_3$　アセチル —C(=O)—CH$_3$　アミノメトキシ H$_2$N—CH$_2$—O—
6	S 24	アルキル基の置換は受けない。その他の原子(団)の置換は受ける。	不可	1価	1	メチル —CH$_3$　エチニル —C≡CH　クロロメチル Cl—CH$_2$—
7	S 30	受ける	可	1価	1	アセトイミドイル —C(=NH)—CH$_3$　アラニル —C(=O)—CH(NH$_2$)—CH$_3$　クロロアセトイミドイル Cl—CH$_2$—C(=NH)—
8	S 32	受ける	可	2価	2	アスパルトイル —C(=O)—CH$_2$—CH(NH$_2$)—C(=O)—　グルタモイル —C(=O)—CH$_2$—CH$_2$—CH(NH$_2$)—C(=O)—
9	S 40	受ける	可	1価	1	フェニル —C$_6$H$_5$　ベンゾイル —C(=O)—C$_6$H$_5$
10	S 41	受ける	不可	1価	1	エチル —CH$_2$—CH$_3$　アリル —CH$_2$—CH=CH$_2$
11	S 42	受ける	可	2価以上	2以上	0-フェニレン C$_6$H$_4$<　エチレン —CH$_2$—CH$_2$—　スクシニル —C(=O)—(CH$_2$)$_2$—C(=O)—　プロパン-1-イル-3-イリデン —CH$_2$—CH$_2$—CH=
12	S 43	受ける	可	2価，3価	1	シクロヘキシリデン C$_6$H$_{10}$<　デカリン-2-イリデン C$_{10}$H$_{16}$<
13	S 44	受ける	不可	2価，3価	1	アリリデン =CH—CH=CH$_2$　エチリデン ≡C—CH$_3$　プロパン-2-イリデン CH$_3$—C(=)—CH$_3$

(3) 官能基

体系名中の主官能基を 下 表に示した。これらも解析・組立てアルゴリズムと密接に関連している。

官能基の種類と分類

項番	分類コード	名称	特徴	例
1	F10	特殊結合官能基	骨格に結合する場合，骨格の1ノードを除去して結合する官能基。 骨格に結合する。	酸  ニトリル −C≡N
2	F20	単純結合官能基	骨格に結合する場合，骨格のノードを変えずに結合する官能基。 骨格に結合する。	カルボン酸  カルボラクトン 
3	F30	アミン類	NH_2R, NHR^1R^2, $NR^1R^2R^3$ の構造となる官能基，骨格，置換基と結合する。	アミン 
4	F40	1原子性官能基	ハロゲン及び擬ハロゲンであり，1原子として挙動する1価の官能基。 骨格，置換基と結合する。	フルオリド −F アジド 
5	F41	カルコゲニド	1原子として挙動する2価の官能基。 骨格，置換基と結合する。	オキシド −O− スルフィド −S−

(4) 特殊機能子

次に示す機能子が，体系名中に出現する。

(a) 結合種の変化を起すもの

① 指示水素，② ヒドロ，デヒドロ

③ Δ（デルタ），④ エン，イン

(b) 原子種が変化するもの

① アザ，オキサなどの類型的命名法の機能子

② シロキサン，シラザンなど

(c) 環系骨格が変化するもの

① ノル，ホモ，セコ，シクロ

② 架橋，③ スピロ化合物

④ 同一要素の集合，ビ，テルなど

(d) 接辞として骨格を置換基化したりイオン化したりするもの

① イル，イリデン，イリジン

② イウム，イリウム

(e) 官能基の二次的修飾

① エステル，② 酸の塩，③ アルコラート，④ 無水物，⑤ ラクトン，ラクタム，⑥ 酸ハロゲン化物，⑦ ケトンの修飾（オキシム，アセタールなど）

1.2 補助構成要素

① ハイフン，② カンマ，③ コロン，④ セミコロン，⑤ ピリオド，⑥ かっこ，⑦ 中点，⑧ スラッシュ

ロカント

体系名中に見られるロカントは，特殊な場合を除いて，おおむね 下 表に示すとおりである。

ロカントの種類

項番	形態	表記例	プライム表記例
1	数字または数字＋英小文字	1, 2b, …	1′, 2′b, …
2	右肩に数字のついた数字	2^2, 7^1, …	$2^{2\prime}$, $7^{1\prime}$, …
3	英小文字	o, m, p だけ	o′, m′, p′
4	ギリシャ小文字	α, β, γ, …	α′, β′, γ′, …
5	英大文字	A, B, AB	———
6	元素記号	N, O, S, …	N′, O′, S′, …
7	右肩記号つき元素記号	N^2, O^2, N^6, …	N^{21}, N^{21}, N^{61}, …
8	数字＋元素記号	1O, 2S, …	1′O, 2′S, …
9	数字＋ハイフン＋元素記号	1−O, 2−O, …	1′−O, 2′−O, 1−O′, 2−O′, …
10	左肩に数字のついた元素記号	^{18}O, ^{14}N, …	$^{18}O'$, $^{14}N'$, …
11	数字（数字）　または 数字(数字＋英小文字)または 数字＋英小文字（数字） 数字＋英小文字（数字＋英小文字）	5 (10), 5 (10a), 5a(10), 5a(10a), …	——
12	数字（数字）英小文字	3 (8)a, …	——
13	英大文字（数字）	A(1), D(15), …	——
14	英大文字（数字＋英小文字） 英大文字（数字＋英小文字，数字＋英小文字，…）	B(9a), AB(10a), … B(9a, 9b)	——
15	数字＋項番10	$1^{10}0$	——
16	同一元素記号が2個連続	OO, T_eT_e	——

1.3 付随情報

これらには，重要なものとして次の3種がある。

① 立体情報

② 同位体標識

^2H, ^3H, ^{18}O など（d，t は使わない）

③ その他の付随情報

旋光方向…(+)(−)(±)

イオン，酸化数…(2+)(2−)のような Ewens-Bassett法か，(II)(III)のような Stock 法を用いる。

体系名中に出現する立体記述子

項番	立体化学の対象		立体記述子
1	キラル点の絶対配置		R, S
2	キラル軸周辺の絶対配置	アレン系	R, S
3		アレン系以外	aR, aS
4	キラル面周辺の絶対配置		pR, pS
5	オレフィン系幾何異性		E, Z
6	環基準面に関する上下方向		α, β, ξ
7	ビシクロ環系の橋		syn, anti
8	糖・アミノ酸の系列		D, L
9	糖の OH の相対配置パターン		arabino, gluco, etc.
10	オレフィン系の相対配置		cis, trans, syn, anti
11	環基準面に関する相対配置		cis, trans, cisoid transoid endo, exo
12	複数のキラル点の相対的配置		R*, S* rel-
13	ラセミ混合物		RS, SR, racemi, rac

2. 体系名の解析と原子結合表の組立て

体系名をプログラムによって解析し，各要素の結合関係を明らかにしつつ，解析辞書中の原子結合表を利用して，化合物全体を組立てる。その処理は，ほぼ次の手順に基づいて行なう。ただし，(5)〜(9)は，体系名のカッコで囲まれた要素群単位に，カッコのレベルの深い部分から，順に繰り返し行なう。

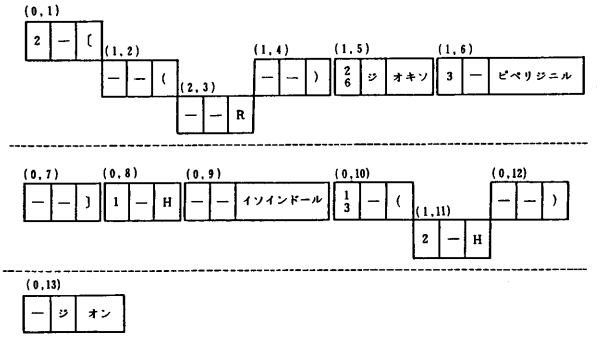
(1) 形態素分析
  ① 解析辞書による各種要素のマッチング
  ② ロカントの認定
(2) 複合語の分割
  （エナール→エン＋アールなど）
(3) そのままでは不明確な構造の明確化
  ① 多重結合位置の決定（ブテン，ヘキセンなど）
  ② 置換基の遊離原子価位置の決定
  （ナフチル，アセナフチルなど）
(4) スピロ構造の形成
  スピロ[4.5]デカン，スピロ[シクロヘキサン-1,1'-[1H]インデン] など
(5) 特殊機能子による処理
  ①ビ，テルなど，②ホモ，③ノル，④セコ，⑤シクロ，⑥類型的命名法による原子置換，⑦指示水素，⑧ヒドロ，⑨デヒドロ，⑩水素，ヒドリド，⑪デ，デス，⑫アンヒドロ，⑬架橋
(6) 接尾辞による処理
  ①エン，イン，②イリデン，③イリジン，④イレン，⑤イウム，イリウム，カチオン，イリオ，イド，イリド，イオン，アニオン，ラジカルなど
(7) 骨格と官能基との結合
(8) 置換基どうしまたは置換基と骨格との結合
(9) 官能基二次修飾子の処理
  （エステル，ラクトン，無水物など）
(10) 原子結合表の該当部分への立体記述子の記入
(11) 全組立て終了時点の正しい結合種の設定
(12) 芳香化，互変異性化
(13) 立体情報に基づいた立体特異的な原子結合表の作成
(14) 拡張モルガン付番と基準環系，主鎖の決定による原子結合表の標準化

次にサリドマイドを例に，原子結合表組立ての手順を示す。

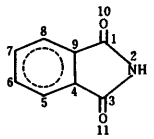
① まず，サリドマイドの入力体系名「2-[(R)-2,6-ジオキソ-3-ピペリジニル]-1H-イソインドール-1,3(2H)-ジオン」の形態素分析を行ない，下図に示すようなレベルに展開する。

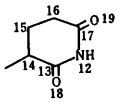
② 次に，ピペリジニルのような手を出す位置が不明な置換基とイソインドールのように指示水素指定を行なう必要のある骨格が存在するので，それぞれピペリジニルの3位から手を出し，イソインドールの1位に指示水素を記入する。すべての要素が明確になった時点で，カッコで囲まれたレベルの深い部分から，結合表の組立てを行なう。

すなわち，イソインドールのような環の場合には，A_1にはある原子からみた左隣りの原子を格納し，A_2には右隣りの原子を格納する。すなわち，ノード1のA_1エリアには9，またA_2エリアには2のノードが記入される。なお，A_3には縮合点のような平面上の，B_1には環の下側の，B_2には環の上側の原子を格納する。



サリドマイドの体系名の形態素分析

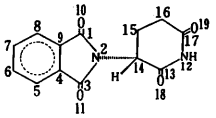




原子の番号	ロカント 1	ロカント 2	原子種	結合価	水素価	立体配置子	結合関係 B1	結合関係 A1	結合関係 A2	結合関係 A3	結合関係 B2
1	1		C	4	0せ			9 1	2 1	10 2	
2	2		N	3せ				1 1	3 1	14 1	
3	3		C	4				2 1	4 1	11 2	
4	3a		C	4				3 1	5 1.3	9 1.3	
5	4		C	3	1			4 1.3	6 1.3		
6	5		C	3	1			5 1.3	7 1.3		
7	6		C	3	1			6 1.3	8 1.3		
8	7		C	3	1			7 1.3	9 1.3		
9	7a		C	4				8 1.3	1 1	4 1.3	
10			O	2						1 2	
11			O	2						3 2	
12	1		N	2	1			17 1	13 1		
13	2		C	4せ	0せ			12 1	14 1	18 2	
14	3	*	C	3せ	1	R	2 1	13 1	15 1		
15	4		C	2	2			14 1	16 1		
16	5		C	2	2			15 1	17 1		
17	6		C	4せ	0せ			16 1	12 1	19 2	
18			O	2						13 2	
19			O	2						17 2	

イソインドール部分　ジオン部分　ピペリジニル部分　ジオキソン部分

標準化前のサリドマイドの原子結合表



サリドマイド(R型)の構造

3. 本システムの特徴

① 日本語でも入出力およびアクセスができる。
② 立体特異的な登録・サーチおよび表示ができる。
③ 体系名だけの入力であるため，構造直接入力よりもコストが1/10ですむ。

4. プログラム

170Kステップ（予想）
CPU 1秒/1化合物で処理

5. おわりに

このシステムは，特徴の項でも述べたように，日英両言語による入出力が可能である点や立体構造を反映した原子結合表を作成することができる点など，従来のシステムにはなかった機能が盛りこまれている。近年，化合物の立体構造がとくに重視されているファインケミカルの分野において，このような化合物データベースが，研究・開発の一助ともなれば，幸いである。また，配位化合物や無機化合物の構造表現など，解決しなければならない問題も残されており，今後も，積み残された化合物タイプの対応と並んで，さらに精密なファイル構築へと，継続していくつもりである。

6. 謝辞

この研究開発は，科学技術庁振興調整費による「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」の一環として，進められているものである。本研究に助力をおしまれなかった「化合物辞書データベース作業分科会」と「JICST 化合物検討会」のメンバー諸氏ならびに特別発表を許可された科学技術庁の関係各位に対し感謝いたします。なお，「作業分科会」のメンバー各位は，下記のとおりです。

〔主査〕
藤原　譲　筑波大学・電子情報工学系
〔委員〕
石塚　忠嗣　㈱日立製作所・ソフトウェア工場
石塚　英弘　図書館情報大学
石原好一郎　㈱住化技術情報センター
大島　輝夫　㈱日本化学物質安全・情報センター
小沢　宏　東京大学・大型計算機センター

酒井　雅英　特許庁・総務部総務課
笹本　光雄　紀伊国屋書店
関根　文三　農薬工業会
竹中　祐典　国立衛生試験所
田島　　真　農林水産省・食品総合研究所
時実　象一　㈱化学情報協会
内藤　裕史　筑波大学・臨床医学系
長谷川正好　帝人株式会社
花井　荘輔　富士写真フイルム株式会社
溝口　次夫　国立公害研究所
山本　　修　工業技術院・化学技術研究所
山本　毅雄　図書館情報大学

参考文献

1）Vanderstouw, G.G. et al.: Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables, J. Chem. Doc. 14(4), p.p. 185～193 (1974)

2）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(I)，第14回情報科学技術研究集会発表論文集，p.p. 101～121 (1977)

3）Choplin, F., Wipke, W.P. et al.: Computer Design of Synthesis in Phosphorus Chemistry: Automatic Treatment of Stereo-chemistry, J. Chem. Inf. Comput. Sci., 18(2), p.p. 110～118 (1978)

4）Morean, G: A Topological Code for Molecular Structures, A Modified Morgan Alegorithm. Nouveau Journal de Chimie, 4(1), p.p. 17～22 (1980)

5）荒木啓介・江種雅子・梶正憲・久米敏雄・平井俊男：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(II)，置換基結合アルゴリズムとその検証，JICST部内報告，1981（昭和56年8月10日付）

6）Dittmar, P.G. et al.: The Chemical Abstracts Service Chemical Registry System (I) General Design, J. Chem. Inf. Comput. Sci. 16(2), p.p. 111～121 (1976)

7）荒木啓介：化合物の立体化学情報記憶方法及び装置，特許公開公報，昭和58－175077，昭和58年10月14日付

8）藤原譲，荒木啓介 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，第21回情報科学技術研究集会発表論文集，p.47－60，(1984)（JICST発行）