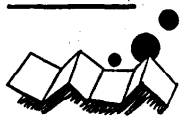
解　説



化学情報のデータベース†

藤　原　譲††

1. はじめに

科学技術の情報のなかで化学の情報は早くからデータベース化され実用に供されてきた．学問的にも産業的にも化学薬品，食品，衣住さらには産業用材料等々の情報は広い範囲からの要求があったことによるものである．情報処理の立場からみて，化学情報も本質的には他分野の情報と同様にデータベースのモデルや理論，また意味論にかかわる基本問題を抱えており，一方では化合物，反応のデータとくにグラフ情報の表現，処理，スペクトル自動解析，物性予測，分子設計などの具体的応用等における特殊な問題も含んでいる．ここでは化学情報の現状，問題点，応用等について概要を述べる．

まず化学情報の必要性と対象について考えると衣，食，住のすべての面で関係する物質や材料は化合物またはその組み合わさったものであるから，生活の面でも生産活動その他の面でも化学の情報抜きでは済まされないのは当然である．また，宇宙開発，海洋開発，航空機，原子力工学などで要求される極限材料や，計算機，マイクロエレクトロニクス，人工臓器，センサなどに必要な特殊機能性材料や，また難病治療用医薬を始め，農薬または環境の立場から高度な生理活性を持つと同時に自然の中での化学的秩序に整合できる薬品等，新しい技術や学問の展開に化合物に関する知識は必須のものである．したがって化学情報のデータベースは直接化学に関係する研究者，技術者のみでなく，関連する自然科学，工学，農学，医学，薬学等の分野の専門家，さらには管理，行政的立場の人，ひいては一般社会人からも程度の差はあれ必要とされるものである．

一方化学の知識やデータは多様であるうえに個別記述的で量が多いため従来より，情報の整理が進み，抄録，データ等が国際的規模で網羅的に収集，利用される体制が確立されており，それらが化学情報のデータベース構築の際に有力な基盤となった．実際にはこれらの情報の従来の媒体であった印刷物の編集の手段として計算機利用が進み，副産物として計算機可読のデータファイルが作成された例が多い．例えば最も代表的な米国化学会で出している抄録誌 CA (Chemical Abstracts) 編集の副産物が CA SEARCH であり，化合物情報を持つ RNSS (Registry of Nomenclature and Structure Service) である．

化合物同定のため赤外分光，可視紫外吸収，質量分析，核磁気共鳴，X線回析等の各種スペクトルの数値データの収集利用も計算機処理に適していたことで，標準スペクトル集積方式が確立していたために早くから実用化され，このことが化学情報の普及に大きな寄与をなした．このうち質量分析のスペクトルデータの自動解析を目的として開発されたシステムが人工知能の初期の成果としてよく知られているスタンフォード (Stanford) 大学の計算機科学，化学，医学の3分野間の共同プロジェクトによる DENDRAL (DENDRitic ALgorithm) であり，MYCIN (感染症診断と抗生物質—MYCIN—の処方システム) への発展の基となった．

化学の情報の特徴の一つは化合物の構造を表現する化学グラフを他の数値および文章データと共に扱うことが要求される点である．このことは2次元，3次元の情報を含み大量で複雑なグラフの処理，識別が持つ問題とともに，データ構造，データ表現，命名，識別番号，標準化等に関する問題を提起している．データベースシステムの立場から見ればデータのモデリング，抽象化，実体の相対性の問題である．

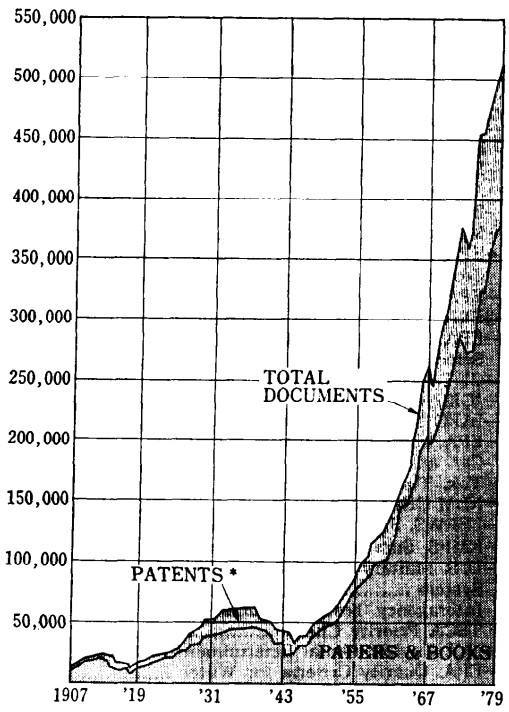
これらの点について以下できるだけ具体的に述べることにする．

2. 化学情報データベースの現状

2.1 CAS のデータベース

化学の分野での代表的データベースは上に述べたよ

† Databases of Chemical Information by Yuzuru FUJIWARA (Inst. of Information Science, The University of Tsukuba).
†† 筑波大学電子・情報工学系



* Includes both patents abstracted and those cited through patent concordance.

図-1 ケミカルアブストラクツ 1907-79年に掲載されている文献数

うに化学関係の世界各国の専門雑誌14,000種を中心に専門の単行本，叢書，特許，学会議事録，技術報告，学位論文までもカバーする CA SEARCH である．図-1 に示すように年間約50万件の文献の書誌事項とキーワード，内容項目等を含み量的にも質的にも非常に高度なものである．CAS (Chemical Abstracts Service) の約1,200名の専従者と各国の文献抄録を含む支援体制に支えられて，広い範囲の情報収集において高度な網羅性と内容の信頼性を維持している．CAS では学会としての各種の出版物編集，印刷にも同じ計算機システムを利用していることから，関係した学術雑誌および書籍の抄録は非常に早く，実際の出版と前後して一般サービス可能となっている．また米国化学会以外の出版物に対しても早期収録に努力している結果，翻訳で時間を要するような特別の文献以外に関しては速報性においてもかなりの水準にある．

CA SEARCH に至るまでに CAS としては文献の題目だけ (Chemical Titles)，および書誌事項とキーワードだけ (CA Condensates) の DB 提供サービスも行っていたが，現在は更に収録内容を増やした CA SEARCH が定着している．また1983年12月より抄録のサービスも始っている[8]．

図-1 はよく知られた Chemical Abstracts に収録されている文献でまさに指数関数的に増加している様子が示されている．これは化学の研究，開発に従事している研究者，技術者の増加ともよく対応しているが，それだけではなく技術の進歩が加速され，各研究者当りの生産性の向上も見逃せない要因である．この傾向は先進国のみに着目すればやがて増加率が低下することが期待できるが世界的にみれば発展途上国の膨大な人口と，抱えている種々の問題点からみて当分情報供給量の増大は続くものと予想される．

CAS で CA 編集における副産物のうちで重要なものの別の代表は，化合物の基本的な情報である化合物名，分子式，構造を示す結合表（約650万件）等を含む RNSS ファイルである．化合物名は索引用に CA でつけた基準名，IUPAC (International Union of Pure and Applied Chemistry) による標準名の他，慣用名，商品名等を含んでいる．構造を正確に表現し，曖昧さがなく (unambiguous 一元対応)，かつ別名もない (unique 一項対応) 体系化された命名法があれば問題はないが，現在化学の代表的国際機関である IUPAC の命名法は一応体系化されているが厳密には一元対応でも一項対応でもないので計算機処理の立場からは充分に体系的とは言えない．一方 IUPAC 名は慣用名とも大きく異なっていることが身近な化合物に多いために人間の利用面からも便利なものとは言えない．結合表は分子内の原子の結合関係のうちトポロジカルな情報を記録する方法としては正確であり，計算機向きでもある．しかし各種の幾何学的，光学的（三次元的）異性体の記述には別の方法が必要とされる．これらは平面グラフであっても生じる問題で，生体関連，高分子，医薬，農薬等では非常に重要な情報で収録対象から除外することはできない．

このような問題を直接解決する方法は避けてデータ管理のために実体 (entity) すなわちこの場合化合物を識別するだけの目的で CAS では登録番号を付与する方法を採用している．これは実用的見地から便利であり，一元対応，一項対応に関しても極く単純な場合は有効であり，現実に CAS 以外で作成されたデータ集，ハンドブック，データベース等でも化合物識別の共通キーとして用いられている．

しかしこのこととはまた後述するように別の深刻な

表-1 SANSS のファイル群

ファイル番号	ファイル名	化学物質数
0	Compounds Not Included In Any Other Collection	13,593
1	EPA. TSCA Inventory List	43,277
2	*CIS. EI Mass Spectrometry	33,898
3	*CIS. Carbon 13 NMR Spectrometry	7,565
4	EPA. Pesticides—Active Ingredients	2,103
5	*CIS/EPA. OHM/TADS	1,005
6	*CIS. Cambridge X-Ray Crystal	14,673
7	Merck Index	8,979
8	EPA. Pesticides—Analytical Ret. Stnds	473
9	EPA. STORET	234
10	EPA. Chemical Spills	577
11	EPA. AEROS SOTDAT	572
12	NIMH. Psychotropic Drugs	2,036
13	EPA. AEROS SAROAD	65
14	*CIS/NBS. Proton Affinities	439
15	CPSC. CHEMRIC	890
16	EPA. Pesticides—Registered Inert Ingredient	779
17	NBS. Gaseous Ions	3,136
18	NFPA. Hazardous Chemicals	395
19	FDA/EPA. Pesticides Ret. Standards	612
21	U. S. International Trade Commission	9,144
22	NBS. X-Ray Crystal	18,169
25	EPA. Elfluent Guidelines	128
26	EPA. Organic Chemical Producers	375
27	IPC. Chemical Product	104
28	IPC. Chemical Plant	103
29	NSF. Chemical List	224
30	EROICA. Thermodynamics	4,488
31	PHS. 149 Carcinogenic Activity	4,416
32	*CIS/NIOSH. RTECS	26,974
33	NIOSH. NOHS	4,556
35	ORNL. EMIC	6,851
36	ORNL. ETIC	3,244
39	EPA/IERL. Non-Criteria Pollutant Emissions	253
40	EPA. Section 111D Clean Air Act	1
42	EPA. Chemical Indicators, Industrial Contamination	49
43	EPA. Selected Organic Air Pollutants	591
45	Clean Air Act—Section 112	5
47	EPA. Carcinogen Assessment Group List of Chemicals	41
53	API/TRC. Thermodynamics and Spectroscopy	3,297
55	WHO/FAO. Pesticide Data Sheets	20
58	EPA/NCTR. Industrial Carcinogen and Mutagen Study	86
59	EPA. Environmental Carcinogen Assessment Program	27
61	DHEW/NCI. Laboratory Chemicals Monographs	55
65	EPA/NSF. Econ. and Tox. Data—Selected Comm. Chem.	499
66	EPA. Restricted Use Pesticides	23
67	EPA. Compounds for Mutagenicity Evaluation	25
68	EPA. Potential Carcinogen Study, Setected Chem.	409
70	CIIT. Priority Chemicals Lists (Toxicological)	27
71	*CIS/JCPDS. Powder Diffraction Patterns	24,130
72	IARC. Monographs (Carcinogenicity Reviews)	388
74	EPA. Expanded Potential Industrial Car. and Mut.	1,444
75	Tox. Tips	472
77	NLM CHEMLINE	31,176
78	USFWS. Fish Control Lab Data Base	91
82	NMFS. Survey Of Trace Elements	15
83	U. S. Military Entomology Information System	123
92	Interagency Testing Committee, TSCA Priority Chem.	4
94	NCI. Environmental Determinants of Cancer	74
95	EPA. Quarity Criteria for Water	37
96	OSHA. Concentration Limits for Gases and Vapors	257
97	NFPA #491 M. Manual of Hazardous Chemical Reactions	899
98	NFPA #325 M. Fire Hazard Properties of Flammables	1,150
100	WHO/FAO. Pesticide Monographs	86
108	EPA. Toxic Substances Control Procedures	21
115	*CIS/EPA. Waler DROP	859
124	OIS. CI Mass Spectrometry	1,147

* のファイルは，CIS で検索可能なデータベース

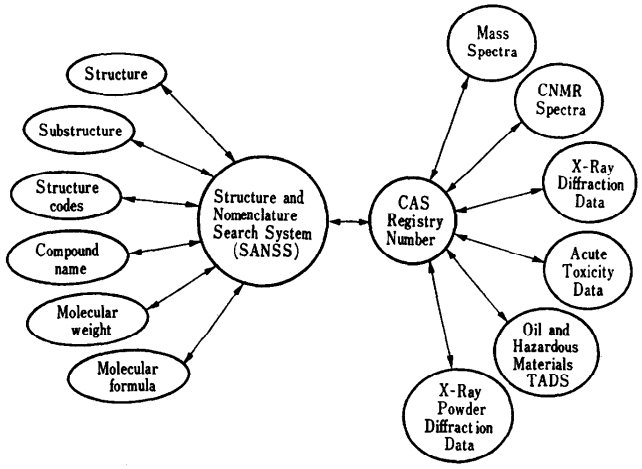


図-2 The NIH-EPA Chemical Information System.

問題を提起することになった.

2.2 ファクトデータベース

化合物の物理的，化学的，生物学的データを最も大規模に集積して早くから一般利用に提供しているのは米国 NIH (国立衛生研究所) の Milne, Feldman および EPA (環境保護局) の Heller 等が開発した CIS (Chemical Information System) で表-1にファイル名一覧，図-2に SANSS (Structure and Nomenclature Search System) を中心に各種データへのアクセス方法を示している．ここで事象データの登録されている化合物の数は全化合物の数に比して極めて少ないことが第一に気づくことであり，またその種類が少なく，結局設計に必要なデータは容易に入手し難いことがわかる.

反応に関しては英国 Derwent 社の CRDS (Chemical Reaction Documentation Service) が国際的に利用されている．特許については同社の WPI (World Patent Index), を提供している．米国の IFI Plenum の CLAIMS, WIPO の INPADOC (International Patent Documentation Center) があり，我が国も特許情報センターの PATOLIS が特許情報のサービスを行っている．特許の場合総称的 (generic), 内包的 (intensional) 表現を用いるので，その問題が未解決で，また網羅性も低い.

表-2 は我が国で作成されているデータベースの代表として日本科学技術情報センターの JICST ファイルを示す．これは科学，技術の分野をカバーしているので化学も含まれている.

2.3 一次情報のデータベース

最近は学術雑誌，ハンドブックの製作に計算機を利用していることから抄録とかデータ抽出を行わないで直接一次情報を入力し，それをユーザに提供する傾向にある.

出版物の内容を直接または二次的にデータベース化して利用することが普及することは，当然データベース中心の雑誌へと展開することになる[4]．最も古いのは New Jersey Institute of Technology の Electronic Information Exchange System であり，また英国でも類似のシステム BLEND (Birmingham and Langhborough Electronic Network Development) が稼動している．これらは生データの獲得や，ワードプロセッサによる文書化，さらに出版のプロセスが計算機を活用していることと，通信技術の進歩による通信コストの低下，高度通信設備の普及，整備の結果として自然な動向と言える.

またオンラインサービスのみでなくより多様なデータ処理に対応するため Bioscience Information Service (BIOSIS) では主題別で索引づきのデータをフロッピディスクで提供しており，版権の問題が残ってはいるが利用の面からは好都合な方式である.

BRS (Bibliographic Retrieval Services Inc.) では Harvard Business Review と Academic American Encyclopedia (Grolier Inc.) の全文検索のサービスを提供しており，さらに Elsevier 版の医学雑誌と ACS 発行の18雑誌について全文検索を1983年6月1日から開始した．ACS関係は1980年以降で1983年1月現在で約30,000件が収録されており，隔週約800件の追加がある.

収録内容は本文，抄録，書誌事項はもちろん，引用文献，脚注，図の説明，CA 登録番号等を含んでいる．ただし，図自体，表，数式等は現在データベース化されていない.

またブリタニカ百科辞典が Mead Data Central でオンライン検索に提供しており，Mc Graw-Hill では31種の雑誌を提供している.

これらの一次情報の直接データベース化はデータ量の飛躍的増加の他に図，表のデータ入力，利用のため新しい機能を必要とする．さらにキーワードの検索でさえ文章内，パラグラフ内，文献内のそれぞれに対応する別々の機能を必要とし，入力データもそれらを識別する構造化が必要となるので DBMS の機能向上が要求されてくる.

このような一次情報の計算機可読化は出版形態の変化と，利用からの要求の他に図書保存の面からも推進

表-2 日本科学技術情報センターで作成し，利用を公開しているデータベース一覧表

	データベース名	収録期間	収録件数	対象分野	内容	サービスの種類
国産データベース	JICST 科学技術文献ファイル	1975年4月～現在	約45万件/年	科学技術全般	JICST発行の「科学技術文献速報」に対応	SDI RS
	JICST 国内医学文献ファイル	1981年4月～現在	約3万件/年	医学・生物科学	医学関連分野の国内誌を対象	SDI RS
	JICST 科学技術研究情報ファイル	1979年～現在	約2万件/年	科学技術全般	日本国内の公共試験研究機関約540機関の研究テーマ	RS

表-3 世界のデータベース作成数
(オンライン/バッチ含む)[6]

	文献データベース	ファクトデータベース	計
1975	335	51	386
1976	337	149	486
1977	422	268	690
1978	533	568	1101
1979	565	715	1280
1980	654	755	1409
1983	762	1083	1845

されている[5]．というのはパルプからの紙を多く使用するようになって100年程であるが，紙の質によって40年ないし100年で紙中の硫酸でセルローズが侵され図書が破損することがわかり，深刻な問題となってきた．化学処理で本の寿命を延す研究も行われ実際にもテストされているが，経費と時間が膨大なものとなる上に昔の和紙と同じ位には寿命が延びないことから，マイクロフィッシュ化と計算機可読化が別の保存策として採用されている．

表-3は従来は計算機可読化データが書誌情報を中心とした，二次情報が主たるものであり，これらも増加しているが，事実の記述である一次情報がより急激に増加していることを示されている．

このことはデータベースの利用がオンライン情報検索した結果を一度利用者によって処理，判断されて次

表-4 化学関係のデータベース

		Food and Agriculture	Forest Products	Textiles	Mining	Metal Products	Environment	Pollution	Energy	Toxicology	Pest Control : Pesticides	Mathematics	Physics	Pharmaceuticals	Geochemistry	Chemical Engineering	Alloys	Polymers	Coordination Compounds	Government Regulations	Patents	Equivalent Patents	CAS Registry Numbers	CA Index Names	Chemical Name Synonyms
CA SEARCH	2,3,4	●	●	●	○	●	●	●	●	●	○	○	●	●	●	●	●	●	●	●		●			●
BIOSIS PREVIEWS	5,55	●	○			●	●		●	●			●												●
NTIS	6	●	○		○		●	●	●	●		○	○	●	●						○				●
COMPENDEX EI	8				●		●	○	●	●					●	●	●								●
AGRICOLA	10	●	●		○		●	●	○	●	●								●						●
INSPEC	12,13										●	●			●										●
FEDERAL INDEX	20	○			○		○	○	○	●	●		●						●						●
EIS PLANTS	22	○	○	○		○				○			○												●
CLAIMS CHEM	23,24	●											●		●	●	●	●		●	●				●
OCEANIC ABS	28	●					●	●	●	●	●			●					●						●
METEOR/GEOASTRO ABS	29						●	●						●											●
CHEMNAME	31										○		●			●	●	●				●	●		●
METADEX	32				●	●									○	●		●							●
WORLD ALUMINUM ABS	33				●	●									○	●									●
SCISEARCH	34,94	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○							●
COMP DISSERT ABS	35	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○							●
ENVIROLINE	40	●			○		●	●	●	●	●		●												●
POLLUTION ABS	41						●	●		○	●														●
PHARM NEWS INDEX	42									●			●						●						●
CA PATENT CONCORDANCE	43																			○	●				
AQUATIC SCI ABS	44	●			○		●	●		●	●			●											●
APTIC	45	○			○		●	●	●	●	●		○						○			●			●
PIRA	48		○						○						○		○								●
CAB	50	●	●				●	●		●	●		●		○										●
GEOARCHIVE	58				●		●	○						●											●
SPIN	62											●	●												●
SSIE	65	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○							●
WORLD TEXTILES	67			●				●									●			○					●
EXERPTA MEDICA	72,73							●		○			●												●
IPA	74												●												●
CONF PAP INDEX	77	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○							●

注) ・主内容 ○副次的内容

2/80

の段階に進むという案内，調査型であったのに対し，直接必要な情報にアクセスし，さらにはその次の処理をも自動化しようとする直接利用，高次処理型になりつつあることを示している．

このことは，データ蓄積，構造，検索機能において大きな影響を与えることになる．

サービス網もよく発達し代表的データベースサービス提供機関である DIALOG に収録されている化学関係のデータベースを表-4 に示す．

3. データベースの応用

3.1 構造活性相関 (QSAR)

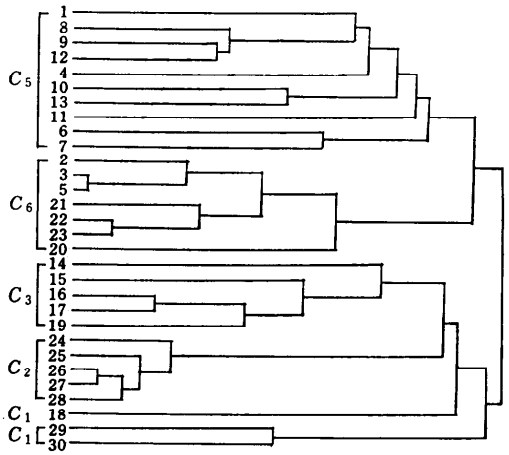
医薬，農薬等生理活性は化学構造と非常に密接な関係があり，最適構造の設計への手掛りは容易に得られそうに見える．実際，この面では各種の試みが行われ成果も得られすでに実用化の段階に入っている．現象的には明確な方向づけができる場合があるが，薬物は投与から各種段階を経由して輸送され代謝過程を経て，特定の場所で生理効果を発揮し，最後に体外に排泄される．これは溶解性，安定性，代謝，薬理活性，分解，長期短期の副作用，環境的影響等のそれぞれの特性がすべて適切な範囲内になければならないことが要求されていることを示す．これらの問題に対しては多変量解析の手法が有効であり，重回帰分析を中心とする方法として Hansch-藤田法や，ADAPT，ALS 法など識別関数を用いる学習機械法が，また分類を中心とするクラスタ分析法などがよく用いられるので主な方法について簡単に説明を行う．

計算機での処理が簡単で結果も明白であるのがクラスタ分析でクラスタリングの基準となる距離のとり方によって特徴に差がでてくるが通常は構造記述因子によるユークリッド距離 d を用いる．

$$d_{ij}=\left\{\sum_{k=1}^{n}(X_k^{(i)}-X_k^{(j)})^2\right\}^{1/2}$$

この距離の近い順に化合物を分類する．図-3 に抗生物質を分類した結果を示す．C_1, C_3 はアミノグリコサイド C_2 はテトラサイクリン，C_5 はペニシリンとセファロスポリン，C_6 はペニシリンとマクロライドから成るクラスタに分かれ菌種の特性を把握するのにも有効であることが示されている[7]．

生理活性の大きさの変化に関与する要因を直接化学構造のみでなく，物理的，化学的性質を用いること，とくにオクタノール/水の分配係数 P の対数である疎水性パラメータ π をはじめ，電子効果 σ，立体効果



Penicillins (1-7)	17 Paromomycin
1 Penicillin-G	18 Gentamicin
2 Methicillin	19 Vistamycin
3 Oxacillin	Macrolides (20-23)
4 Ampicillin	20 Erythromycin
5 Cloxacillin	21 Spiramycin
6 Sulbenicillin	22 Kitasamycin
7 Carbenicillin	23 Oleandomycin
Cephalosporins (8-13)	Tetracyclines (24-28)
8 Cephalothin	24 Chlorotetracycline
9 Cephaloridine	25 Oxytetracycline
10 Cephaloglycin	26 Tetracycline (1)注)
11 Cephoxitin	27 Tetracycline (2)注)
12 Cephazolin	28 Pyrrolidinomethyl-tetracycline
13 Cephalexin	
Aminoglycosides (14-19)	Peptides (29, 30)
14 Streptomycin	29 Colistin
15 Neomycin	30 Polymyxin B
16 Kanamycin	

注) 測定日が異なる．

図-3 抗生物質のデンドログラム

Es などを要因の記述子として用いるのが Hansch-藤田法である．いま一定の効果を示すのに必要な活性物質の溶度あるいは薬量を C として，次の関係式を用いて重回帰分析を行う．

$$\log\frac{1}{C}=a\pi+\rho\sigma+\delta Es+\cdots+\text{constant}$$

各要因は線型のものが多いが，疎水性パラメータは二乗の項まで入っているのがこの方法の大きな特徴である．

図-4 に示すように活性は $\log P$ に対し放物線を画き最適値の存在が明確になる．とくにこの例ではイネとヒエのように類似した植物に対する有効な除草剤の開発といった，高度な選択性の要求されるような場合にも役立つことを示している[8),26)]．

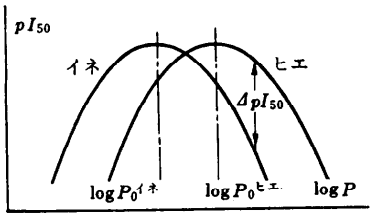


図-4 除草活性の疎水性に対する関係の植物種間における相違

構造活性相関における構造的要因をできるだけ機械的に処理する方法として前節で述べた重回帰分析に対して，パターン分類すなわち識別関数を用いる方法はパターン認識法，または識別関数を訓練用データを用いて学習することから学習機械法とも呼ばれる．森口の開発した適応最小二乗法 (Adaptive Least Square Method-ALS) は通常の識別関数が対象データを2分するのに対し，ALS 法では一段で対象を3以上のグループに分類予測する点に大きな特長がある．図-5にその概要を示すように識別関数は各構造要因と重みベクトルとの内積で他の学習機械方式と同じであるが，m個のグループに対し $(m-1)$ 個の分割境界値を持っており，この境界値は対象データと分類の数に適応するように定められる[9]．

ALS 法は計算機の使用を前提としているので要因の数を人間のばあいには扱えない程大きく，例えば50～100 またはそれ以上にすることが容易であり，また定性的記述と定量的記述の両方が混在していても解析できること，識別関数も直接薬物の活性を定量的に示すものでなくて，定性的表現であっても良いこと，

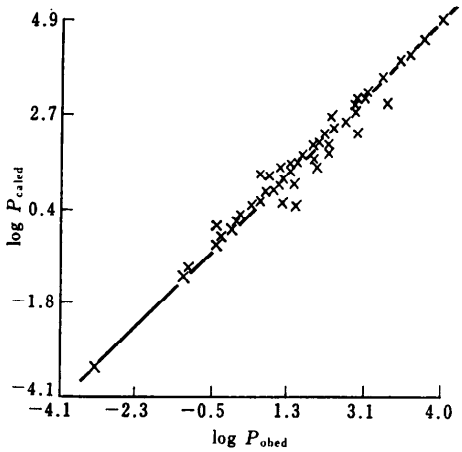


図-6 76のテスト化合物について $\log P_{obsd}$ 対 $\log P_{calcd}$

さらに各要因の活性への寄与の仕方が全体として把握できることなどから多くの例について試みられその有効性が認められている．

パターン認識法としては Jurs の ADAPT (Automated Data Analysis Using Pattern Recognition Techiques) が有名である．基本的には ALS 法と同じ線型識別関数を用いるが，このシステムは化学，薬学の研究者向きに各種の便利な機能を備えている．一つは Hansch 法で重要な役割を持つ疎水性パラメータが必ずしも他のデータと共に与えられるとは限らないので，$\log P$ の推算プログラムが組み込まれている点である．図-6 は $\log P$ の推算値と実測値を比較したもので，良く一致しているのがわかる[10]．

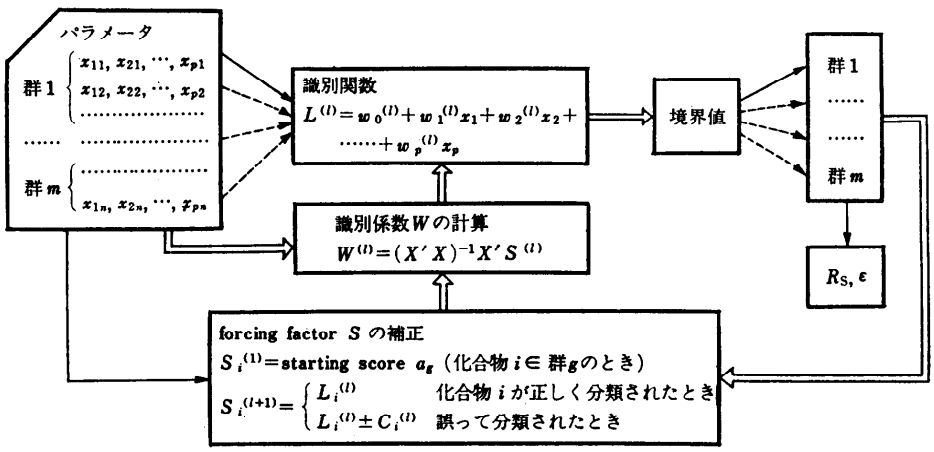


図-5 ALS の手順

3.2 反応設計システム

化学においては反応が中心課題であるが，これまでに述べてきた多様性が反応の場合最も顕著に現われる．したがって合成経路の設計は極めて複雑になる反面，計算機による反応設計への期待がそれだけ大きくなっている．最適の合成経路解析のためのシステムは構造活性相関と同程度に早くから関心が持たれており，1969年に有名な Coreyのシステム LHASA (Logic and Heuristics Applied to Synthetic Analiysis) の前身 OCSS (Organic Chemical Simulation of Synthesis) が報告されている[11)~13)].

LHASA を開発した Corey の共同研究者の一人 Wipke はデータ入力を柔軟にし立体化学を組み込んだ SECS (Simulation and Evaluation of Chemical Synthesis) を開発した．これはさらに REACCS (Reaction Access System) と呼ばれる市販システムとして利用できる段階になっている．これらはプログレマンシステムの方式になっており，図-7 に示すように目標化合物を合成する前駆物質を指定し，それぞれの前駆物質を更に前にさかのぼる．各反応の逆向き変換を transform，合成の逆を antithesis と称し，目標化合物から原料まで合成経路が遡れることが反応の解析である．各段階で考えられる前駆物質が CRT 上に表示される．この際逆反応の計算機内部処理は結合の切断，生成，原子の除去，附加などであり，網羅的に行うことは容易である．これを次々にそのまま繰り返すと膨大な組み合わせが出てくるので各段階で研究者と対話しながら選択を行って次へ進む．この際反応をその構造的特徴に従って点数を与えその点数を増減して得られた評点が選択の一つの基準になる．評点がある点 (−50) 以下では，自動的にその反応は対象から外される．Corey はこの分野の開拓者的存在であると共に反応の計算機処理のために必要な概念を定め，また反応を計算機向きに整理したことでも大きな功績を残した．

反応設計システムとしては Bersohn によるバッチ型[14)], Gelernter によるバッチ-TSS 併用型[15)], もある．

米田は反応系と生成系をそれぞれ結合行列で表現し，それらの差として反応行列を定義した[16),17)]．反応系には原料化合物のみならず触媒も含められるので触媒を含めた系の解析が行えることと，主反応とともに副反応までの動力学が扱えることを特長とするシステム GRACE (Generation of Elementary Reaction Network in Radical Reactoins) であり，工業生産の場での反応解析に適したシステムである．図-8 に例

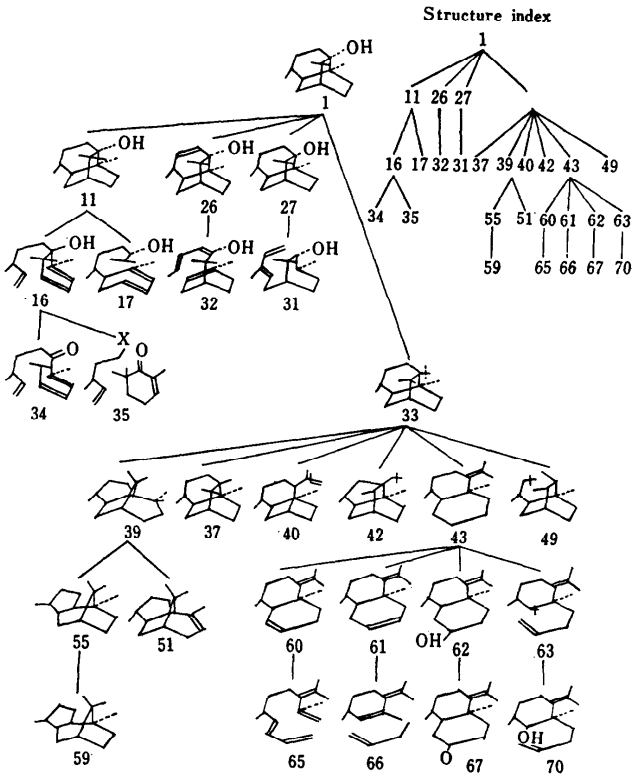


図-7　コンピュータによる合成経路解析の例

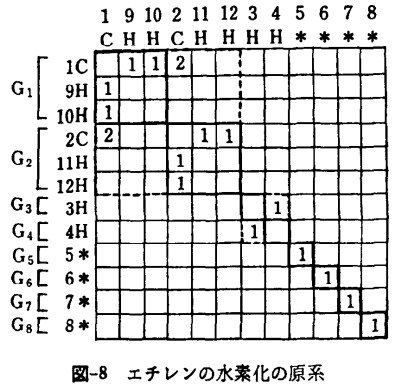


図-8　エチレンの水素化の原系

を示す．反応行列を用いることは計算機向きと言えるので Ugi も同じ方式を採用して CICLOPS (Computers in Chemistry, Logic Oriented Planning of Synthesis) システムを開発した．

反応行列式は反応の機械的処理とファイル化に適しているが，体系化をより意識したシステムが Hendrickson により開発されている[28]．

4. データベースおよびデータ処理の問題点

化学のデータを整理するには化合物の体系的表現が必要である．データへのアクセスが化合物の構造に対応したキーになっていれば自然であり，便利でもある．

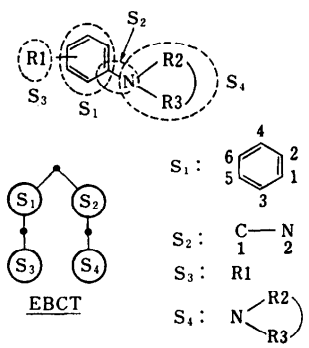
上で述べたように RNSS では，650万を超す化合物が登録されており，データ項目としては化合物の標準名，慣用名，分子式，結合表による構造情報，図形出力用情報，立体化学，同位体等のデータも含まれている．また種類としては確定構造を持つ有機の純物質が全データの約80%を占めており，無機化合物，錯体，高分子，金属，合金，混合物，構造不明物質，総称的記述による一連の化合物も収録されている．ただ有機の純物質は発表されたすべての化合物が登録されているが，それ以外のものは識別が明確な形でかつ，実際に合成または分離されたものをデータの収録対象とするので，必ずしも網羅的ではない．したがって特許，材料，混合物等では利用の際に注意しなければならない．CAS の最大の特長は化合物に登録番号(CAS Registry Number)を与えており化合物のキーとして一項対応 (unique) でかつ一元対応 (unambiguous) の識別システムを目指して努力が積まれており，CAS 以外のデータベース構築の際にも利用されている．

我が国においてはデータベース構築は種々の事情で先進国の中では例外的に遅れをとっていたが昭和56年より化合物総合データベースシステムが国家的プロジェクトとして5年計画で動きだしたことで今後の見通しが得られるようになってきた（表-5 参照）．

設計のような研究開発支援システムにおいてはデータの表現とくに知識システムで処理できるためには物性制御，機能開発にかかわる構造要因の記述，表現が重要である．一般に現在のデータベースが個別的，外延的データから構成されているが，総称的，内包的データを収録する必要がある．とくに特許情報の Markush 表現，法律用語などは実用的に非常に大きな問題がある．APL II の一般行列 (general array) を利用した一つの解決法の例を図-9 に示す．

表-5 総合化合物データベース

データベース名	略称	作成機関	スコープ
化合物辞書データベース	DI	日本科学技術情報センター(社)化学情報協会	構成各データベースに収録する化学物質を登録し，それらの名称・構造・所在等の情報を収録
バイオケミカルデータベース	BC	農林水産省食品総合研究所	肥料・食品添加物・酵素等，生化学的応用をもつ物質に関する諸データを収録
バイオロジカルデータベース	BL	国立衛生試験所	変異原性・小核誘発性・催奇形性等，化学物質が生物に及ぼす活性・毒性に関するデータを収録
環境データベース	EN	国立公害研究所	環境調査対象物質について，環境測定データや分析データを中心に必要データを収録
医薬品データベース	PH	(社)日本医薬情報センター	医薬品の種々の名称（商品名・慣用名・略名等），記号ないし番号，取扱業者等の情報を収録
安全性データベース	SF	(社)日本化学物質安全・情報センター	化学物質について規定した国内の法規制についての情報を収録
スペクトルデータベース	SP	工業技術院化学技術研究所	基礎的な標準物質の赤外，^{1}H-NMR・^{13}C-NMR の各スペクトルデータを収録
熱物性データベース	TH	日本科学技術情報センター	一成分系及び二成分系の熱力学的・熱化学的性質を中心とした物性データを収録
農薬データベース	PE	農薬工業会	農薬の物性，安全性，用途，使用法等に関するデータを収録
急性毒性データベース	TX	(財)日本医薬情報センター	中毒に関係する物質，生活関連物質の毒性，症状，治療法等に関するデータを収録



	S1	S2	S3	S4
S1	0	1	b(2,4,6)	0
S2	1	0	0	s(2)
S3	b	0	0	0
S4	0	s	0	0

ECM

図-9 総称構造の EBCT 表現

また正確かつ柔軟なデータへのアクセスはデータ利用上重要であるが前節で述べたCASの登録番号のようにキーとして疑問の余地のない程確立し，他に適当な代替案がなさそうに見えることでも表-6に示すように混合物，総称表現，または技術の進歩に伴う化合物識別レベルの変化による更新などの理由により，便利または正確な利用が容易とはいえない状態になっている．

表-6　化合物登録番号の例

	名　称	登録番号
元素・イオン	Aluminium	7429—90—5
	Al^{1-}	22325—47—9
	Al^{1+}	14903—36—7
	Al^{2+}	15724—40—0
	Al^{3+}	22537—23—1
	Al^{4+}	14700—92—6
	Al^{5+}	16997—96—9
	Al_2^{1-}	52932—00—0
	Al_2^{1+}	12595—57—2
	Al_3^{1+}	12595—58—3
同位体	Al—24	15714—99—5
	Al—25	15756—00—0
	Al—26	14682—66—7
	Al—28	14999—04—3
	Al—29	14867—31—3
合金	Al 60-65, Nd 35-40	74989—11—0
	Al 88-97, Nd 2.6-12	74129—55—8
	Al 88-100, Nd 0-12	73965—57—8
無機化合物	一酸化炭素	630—08—0
	^{11}CO	10456—04—9
	^{13}CO	1641—69—6
	$^{13}CO^{1+}$	74520—33—5
	$^{13}C^{17}O$	53334—35—3
	$^{13}C^{18}O$	35907—63—2
	^{14}CO	7665—54—5
分子化合物	Chinoform	130—26—7
	Chinoform glucuronide	34296—97—4
	Chinoform Mg chelate	43019—17—6
	Chinoform sulfate	16524—52—6
有機塩	L-Ascorbic acid	50—81—7
	L-Ascorbic acid monosodium salt	134—03—2
	L-Ascorbic acid sodium salt	7317—67—1
重合体	Polyvinyl chloride	9002—86—2
	Polyvinyl chloride isotactic	26793—37—3
	Polyvinyl chloride syndiotactic	25037—47—2
混合物	L-Ascorbic acid -$FeSO_4$ (1 : 1)	55128—83—1
	-$FeSO_4$-$NaHCO_3$	39284—32—6
	-Vitamin B_{12}	39389—85—0
	-Vitamin B_1-L-Cysteine	56333—08—5
	-Vitamin A-Vitamin E	39298—16—3

日本科学技術情報センター提供資料

このことは情報処理の技術の向上すなわちより速く，より大きく，より正確なシステムの実現以上にデータモデルを中心とするデータベース理論の展開が必要なことを示唆している．

またデータの集積自体もそれぞれの分野の専門家の多くの時間と知識を必要とすることが，国際的にも，我が国においても深刻な問題となっている．これは関係者の努力もさることながら新しい分野の育成と種々の制度上，意識上の障害の低減も必要である．

化学情報の中心は化合物の構造で行列，またはそれに対応した結合表によって計算機内で表現される．部分構造も含めた化合物間の構造の比較は標準名，慣用名，結合表のいずれを用いても精度，速度，柔軟性で不充分であり，登録番号によっても充分な解決が得られていない．

金属，無機化合物，高分子などではこのことが特に顕著に現われている．

頻繁に使われる化合物に対してはシノニムの数が非常に多く，ポリスチレンの場合300以上，ポリエチレンに至っては現在1,000を超えるシノニムがある．その一部を表-7に示した．これはMOLECULEを使

表-7　200以上の同意語をもつ化合物

No. of Synonyms	Reg. No.	Compound name
1027	9002-88-4	Ethylene, polymers, (C 2 H 4)×
708	9002-86-2	Ethylene, chloro-, polymers, (C 2 H 3 C 1)×
418	9003-53-6	Styrene, polymers, (C 8 H 8)×
390	9003-07-0	Propene, polymers, (C 3 H 6)×
316	9003-35-4	Phenol, polymer with for maldehyde,(C 6 H 60. CH 20)×
299	24937-78-8	Acetic acid vinyl ester, polymer with ethylene, (C4 H 602. C2 H 4)×
266	9016-45-9	Glycols, polyethylene, mono (nonylphenyl) ether, (C 2 H 40) nC 15 H 240
257	147-14-8	Copper, (phthalocyaninato (2-))-, C 32 H 16 CuN 8
244	25068-38-6	phenol, 4, 4′-isopropylidenedi-, polymer with 1-chloro-2, 3-epoxypropane, (C 15 H 1602. C 3 H 5 C1 0)×
243	9003-56-9	Acrylonitrile, polymer with 1, 3-butadiene and styrene, (C 8 H 8. C 4 H 6. C 3 H 3 N)×
238	9003-08-1	Melamine, polymer with formaldehyde, (C 3 H 6 N 6. CH 20)×
203	25038-54-4	Poly (iminocarbonylpentamethylene), (C 6 H 11 NO)n

表-8 JOIS 検索補助用データベース

データベース名	内 容
JICST科学技術用語シソーラスファイル	JICSTファイルに使用されている統制語キーワードが収録されている．JICSTファイルを検索する時使用する．
JICST 資料所蔵目録ファイル	JICSTファイルを資料番号から検索する時資料番号を入力する必要があるが，このファイルによって雑誌名からこの資料番号がわかる．
CA SEARCH 化合物ファイル	毎週発行される CA SEARCH ファイルから JICST が化合物名の巻末索引語部分だけをぬきとって作成するものである．したがってこのファイル中の化合物名は CAS の正式名称のみからなり，同義語は含まれていない．
MeSH 医学用語ファイル	NLM が作成するもので，MEDLINE ファイルに使用されている統制語キーワードが収録されている．MEDLINEファイルを検索する時使用する．

って出したもので[27]，システムによって数は異なるが，名前から検索する場合，これだけのものを入れなければ万全は期し難い．もちろん，これだけのシノニムを全部人間が憶えておいて検索の質問式を作るのは不可能なことだが，コンピュータでも容易ではない．これらに対し専門家の膨大な人手を要するシソーラスが有効であることはよく知られており，表-8 にその例を示す．これはユーザには便利であるが維持が大きな問題である．

これらはデータの意味論的取り扱い[23]，抽象化[24),25]を必要とするが現実のデータに対しては充分ではなく減別表現を含み多くの未解決問題が残されている．

グラフィックスの有効性についてはデータの特性上当然であるがグラフィックスのためのデータベース構造，処理言語体系，グラフ表示の標準化，高度処理のアルゴリズムとハードウェア等の問題が残されている[21),22]．

5. おわりに

化学の研究においては日常の研究活動の多くの段階すなわち調査，検索で文献データベースを用い，装置制御，測定データ処理から合成設計，薬品設計等にファクトデータベースを活用した研究方法が定着し，さらに高度な知識活動支援への展開が期待されている．一方計算機はより手軽により高性能になりつつあるのであるからこの面からもデータベースを利用した化学研究の効率化が促進されるであろう．またこのことはデータ表現，データ構造，知識表現と処理等に問題を提起することにもなる．

参考文献

1) CAS Today 2 (1980).
2) Heller, S. R. et al.: Anal. Chim. Acta/CTO, pp. 117-122 (1980).
3) CAS Report p. 13 (Nov. 13, 1983).
4) Seldom, W. Terrant: Chem. and Eng. News, p. 51 (Apr. 25, 1983).
5) LC Preservation Leaflet No. 1 (Mar. 1982).
6) EUSIDIC Database Guide (1983).
7) Takahashi, Y. et al.: Anal. Chim Acta/CTO, pp. 122-241 (1980).
8) Fujinami, A. et al.: Pestic. Biochem. Pgysiol., pp. 6-287 (1976).
9) Moriguchi, I. et al.: Chem. Pharm. Bull., pp. 25-2800 Tokyo (1977).
10) Chou, J. T. et al.: J. Chem. Inf. Comput. Sci., pp. 19-172 (1979).
11) Corey, E. J. et al.: Science, pp. 166-178 (1969).
12) Corey, E. J. et al.: J. Am. Chem. Soc., pp. 94-421 (1972).
13) Wipke, W. T. et al.: Computer Graphics, pp. 5-10 (1971).
14) Bersohn, M. et al.: J. Chem. Inf. Comput. Sci., pp. 19-137 (1979).
15) Gelernter, H. L. et al.: Comput, and Chem., pp. 2-75 (1978).
16) 米田幸夫：ケモグラム，丸善 (1972).
17) Yoneda, Y.: Bull. Chem. Soc. Japan, pp. 8-52 (1979).
18) Nakayama, T. and Fujiwara, Y.: J. Chem. Inf. Comput Sci., pp. 23-80 (1980).
19) Fujiwara, Y. et al.: Data for Sci. and Tech. (CODATA) pp. 150-153 (1983).
20) 藤原 譲：分子設計研究会議事録 No. 204 (1982).
21) Dubois, J. E. et al.: Compt. Rand. Acad. Sci. Paris, p. 292-783 (1981).
22) Dubois, J. E. et al.: Bull. Soc. Chim. Fr., p. 1390 (1975).
23) Codd, E. F.: Extending the Database Relational Model to Capture More Meaning, ACM TODS 4 (4), pp. 397-434 (1979).
24) Smith, J. M. and Smith, D. C. P.: Database Abstraction: Aggregation, Comm, ACM 20(6), pp. 405-413 (1977).
25) Smith, J. M. and Smith, D. C. P.: Database Abstraction; Aggregation and Generalization, ACM TODS 2 (2), pp. 105-133 (1977).
26) Hansch, C. et al.: J. Am. Chem. Soc., pp. 86-1616 (1964).
27) Ozawa, H., Ishizuka, H. and Chihara H.: V Symp. Inf. Chem. (Dec. 1983).
28) Hendrickson, J. B.: J. Chem. Inf. Comput. Sci., pp. 19-129 (1979).

(昭和59年3月22日受付)