In Silico Data
Miracles by the KY-methods

The Chem-Bio Informatics Society 情報計算化学生物学会

CBI学会2019年大会、2019年10月21日（月）、13:00－17:00

〈チュートリアル〉

計算毒性学と化学データサイエンスの基本

株式会社　インシリコデータ
湯田　浩太郎

本日のプログラム：

1．13:00-13:05:（5分） 挨拶：株式会社　インシリコデータ　湯田浩太郎

2．13:05-13:20（15分） ◆導入　計算毒性学と「化学データサイエンス」
　計算毒性学でのコンピューター導入原理、二大毒性評価関連技術（化学多変量解析／パターン認識アプローチ、人工知能アプローチ）、データサイエンスから「化学データサイエンス」へ

3．13:20-13:50（30分） ◇第一部　計算機化学（Computer Chemistry）関連
　化合物保存形式、化合物命名法、化合物検索（完全一致、部分構造、2・3次元構造検索、他）手法、一元一項対応串刺し検索、化合物の扱い（縮合多環、互変異性、立体／幾何異性）、化合物表記（ケトエノール、ニトロニトロソ、他）

4．13:50-15:20（90分） ◇第二部　化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連
　化学パラメーター、2／3次元パラメーター、種々データ解析手法、過剰適合、偶然相関、線形／非線形性、特徴抽出、最少サンプル数、最少パラメーター数、クラスポピュレーション、次元変換／圧縮／縮小、分類率／予測率、要因解析、オートスケーリング、アウトライヤー／インライヤー、解析信頼性指標（サンプル数／パラメーター数）、KY（K-step Yard sampling）法、パーセプトロン、バックプロパゲーション、遺伝的アルゴリズム、ファジー理論、内挿／外挿問題、他

＜15:20-15:40　休憩 20分＞

5．15:40-16:20（40分） ◇第三部　人工知能（Artificial Intelligence）関連
　人工知能の歴史、ルールベース型人工知能、ニューラルネットワーク型人工知能、深層学習、サンプル数問題、要因説明問題、ルールのコンピューターへの組み込み、ネットワーク構造、LISP、FORTRAN、PYTHON、

6．16:50-17:00（30分） ◇第四部　計算機科学（Computer Science）関連
　データベース理論、プログラミング言語、クラスター、クラウド、スーパーコンピューター、ネットワーク、WEB、他

7．16:50-17:00（10分） ◇討論および名刺交換会

オブジェクト指向
ネットワークの基礎
並列プログラミング

1.基礎
2.計算量
3.戦略
4.データ
5.アルゴリズム
6.データベース
7.コンピューター
8.プログラミング
9.おわりに

戦略

反復処理
再帰処理
総当たり攻撃
バックトラック戦略
発見的解法
分割統治法
動的計画法
分岐限定法

データ

基本データ型
代表的な抽象データ型(ADT)が取り上げられています。
スタック、キュー、優先度付きキュー、リスト、集合済みリスト、ソート済みリスト、マップ、集合。
データ構造
配列、連結リスト、双方向連結リスト、木、二分探索木、二分ヒープ、グラフ、ハッシュ表。

アルゴリズム

ソート
選択ソート、挿入ソート、マージソート、クイックソート。
探索
線形探索、二分探索。
グラフ
深さ優先探索、幅優先探索、経路探索。有名なダイクストラアルゴリズムや、Googleのページランクアルゴリズムが取り上げられています。
オペレーションズリサーチ

基礎

数論
組合せ数学
グラフ理論
数理論理学
情報理論
理論計算機科学
　計算理論
　チューリングマシン
　ラムダ計算
　アルゴリズム解析
データ構造
形式言語
プログラム意味論
　領域理論
計算科学

6. 化合物関連データベース

■データベースのコンテンツによる分類

- 特許データベース
- 化学文献データベース
- 化合物物性データベース
- 化合物機器スペクトルデータベース
- X線結晶データベース
- ポリマー関連データベース
- 蛋白関連データベース
- バイオ関連データベース
- 薬理活性関連データベース
- 毒性関連データベース
- 医療関連データベース
- その他

6. 化合物関連データベース

◇特許データベース

データベース名	収録	収録期間	特徴
商用データベース（日本特許）			
PATOLIS IV	パトリス	書誌1955- フリーキーワード1971-	日本特許DBの中では信頼性が高い。 1989年からフルテキストの全文検索も可能に。
JP-NET	日本パテントデータサービス	1989-	検索ワードのハイライト機能、審査経過情報へのリンク。
NRIサイバーパテントデスク	NRIサイバーパテント	1971-	1971年以降の全文検索、概念検索も可。
SRPARTNER	日立情報システムズ	1983-	概念検索、近接演算可能。
RIPWAY	テコーテクノシステムズ	1983-	概念検索も可。分析機能あり。
CKS-Web	中央光学出版	1993-	SDIに重点。
DocuPat	富士ゼロックス	1978-	1978年以降の全文検索可。
HYPATWEB	発明通信社	1983-	1983年以降の全文検索が可。検索精度は良好。
公開技報	発明協会	目次1992-、全情報2002-	SDIサービスも。
商用データベース（外国特許）			
CLAIMS	US	citation 1947- legal status 1980-	引用特許 法的状況
INPADOC	WorldWide	1968-	66の国・機関の対応特許、22の国・機関の法的状況
DPCI		1973-	引用特許情報
MARPAT		1961-	CA収録特許のマルクッシュ構造特許
DWPI	WorldWide	1963-	抄録検索、対応特許情報の他、各種の独自コード類での検索
DII(Derwent Innovations Index)	WorldWide	1963-	抄録検索、対応特許情報の他、一部コードを除く独自コード類で検索
QPAT	WoldWide	US 1971-,EP 1978- PCT 1978-	US,EP,PCT全文検索の他、DWPIと同様WorldWideな検索も可能。 1特許1レコードのPlusPatと1ファミリー1レコードのFamPatがある。
Delphion	US,EP,PCT,DE	US 1971-,EP 1986- PCT 1978-	US,EP,PCT,DE全文検索の他、DWPIも別課金で、検索できる。 統計分析などの付加機能も。
PatentWEB	US,EP,PCT全文 GB,DE,FR,JP抄録	US 1836-,EP 1978-	US,EP,PCT全文検索の他、GB,DE,FR,JPの抄録、INPADOCの検索が可。 1IDで複数同時アクセスができる。
FOCUST	US,EP,PCT	US 1976-,EP 1978-	US,EP,PCT全文検索。引用情報などの解析機能が豊富。
PatBase	US,EP,PCT,GB,DE	US 1976-,EP 1978-	US,EP,PCT,GB,DE,FR全文検索の他、CN,KRの抄録検索も可。
WIPS	US,EP,PCT	US 1976-,EP 1978- PCT 1978-	US,EP,PCT全文検索の他、KR,CNの抄録検索も可。 KRの収録はタイムラグが商用DB中で最も小さい。
PATLIST-CN	CN	1985-	日本語でも検索できる中国特許DB
WORLDPATENT.JP	CN(KR予定)	1985-	日本語入力で中国語検索キーワード候補ー検索
無料データベース			
IPDL テキスト検索	JP	テキスト検索 1993-	要約とクレームからテキスト検索が可能。各公報から審査経過情報へも直リンク 公報のイメージデータは、明治18年～画像あり
USPTO	US	登録 1790- 公開 2001-	米国特許全文(フルテキスト)の検索が可能。 審査経過・包袋情報はPAIRのページで検索可。
espacenet	WorldWide	EP 1978-,US 1968-, PCT 1997-	検索結果から図面表示、審査経過情報(INPADOC)の表示も可。 審査経過・包袋情報はepoline Register Plusで。
WIPO	PCT	1978-	全文検索は1996-。各国への移行情報、サーチレポート（収録は不十分）。
KIPRIS	KR	1948-	韓国特許情報院(KIPI)が運営。タイムラグなしで審査経過情報が得られる。
中国特許庁	CN	1985-	公開・登録の公報表示可。審査経過情報は別ウィンドウで検索可。
CNIPR	CN	1985-	中国知識産権網。全文検索、コマンド検索可。
台湾特許庁	TW	1950-	公報表示、審査経過情報などが入手できるが、検索項目は少ない。
TWPAT	TW	書誌 1950- クレーム 1974-	検索・抄録表示は無料。コマンド検索可。 公報表示、審査経過情報、ダウンロード機能などは有料。
中華民国全国工業総会	TW	1985-	無料では5件までクレーム表示。

http://www.geocities.jp/patentsearch2006/DB-table.html

6. 化合物関連データベース

◇文献データベース

ﾃﾞｰﾀﾍﾞｰｽ名	提供元	収録年	収録件数	収録内容
JSTPlus JDream Ⅱ	JST	1981-	約1300万件	科学技術（医学を含む）全分野に関する文献情報
JST7580	JST	1975-1980	約220万件	科学技術文献情報
JMEDPlus	JST	1981-	約240万件	日本国内の医学文献
CAplus	STN	1840-	25900000	化学、化学工学の文献と特許（対応特許も収録）
CAold	STN	1907-1966	2600000	CAの書誌、CAS登録番号を収録
PASCAL	Dialog	1973-		科学技術文献
RAPRA	STN/Dialog	1972-	894000	ｺﾞﾑ、ﾌﾟﾗｽﾁｯｸ、接着剤、高分子関係の文献、特許
COMPENDEX	STN	1970-	68500000	工学、技術に関する網羅的な文献
INSPEC	STN	1969-	8900000	物理学、電気工学、ｴﾚｸﾄﾛﾆｸｽ、ｺﾝﾋﾟｭｰﾀ関連文献
WSCA	STN	1976-	284000	塗料、表面処理、添加剤、接着剤関係文献
PIRA	STN	1975-	456000	包装、印刷、紙などの技術開発に関する文献
GeNii CiNii	国立情報学研究所			文献検索
JOPAL	WIPO	1981-		文献がIPCでも検索できる(収録ｼﾞｬｰﾅﾙ一覧)
CSDB	JPO			ｺﾝﾋﾟｭｰﾀｿﾌﾄｳｪｱﾃﾞｰﾀﾍﾞｰｽ(収録範囲)
combined chemical dictionary	CHEMnetBASE			有機化合物辞典など５種類の辞典の統合版
ﾊﾞｲｵ関係ﾃﾞｰﾀﾍﾞｰｽ				
BIOSIS	STN	1969-	15600000	生物、生物医学文献
EMBASE	STN/Dialog	1974-	10900000	生物医学、薬学文献
MEDLINE	STN/Dialog	1950-	15800000	歯科、看護学、環境科学など生物医学、薬学文献
DGENE	STN	1981-	7620000	特許中の核酸配列、ﾀﾝﾊﾟｸ質のｱﾐﾉ酸配列
PDBｱﾐﾉ酸配列検索	産総研 生命情報科学技術ｾﾝﾀｰ			ｱﾐﾉ酸配列検索
PDBj	大阪大学蛋白質研究所)			蛋白質構造ﾃﾞｰﾀﾊﾞﾝｸ
RCSB PDB				蛋白質ﾃﾞｰﾀﾍﾞｰｽ
SWISS-PROT	ExPASy Proteomics			蛋白質のｱﾐﾉ酸配列のﾃﾞｰﾀﾍﾞｰｽ
文献、核酸・蛋白質	米国国立ﾊﾞｲｵﾃｸﾉﾛｼﾞｰ ｲﾝﾌｫﾒｰｼｮﾝｾﾝﾀｰ(NCBI)			文献情報、核酸・蛋白質などのﾃﾞｰﾀﾍﾞｰｽ
PubMed	NCBI			文献検索ﾃﾞｰﾀﾍﾞｰｽ
OMIM遺伝子ﾃﾞｰﾀﾍﾞｰｽ	NCBI			遺伝子の概要を把握できる
iyakuSearch	日本医薬情報ｾﾝﾀｰ			国内外の医薬品情報ﾃﾞｰﾀﾍﾞｰｽ 医薬文献情報検索
生物遺伝資源	製品評価技術基盤機構			ﾎﾓﾛｼﾞｰ検索

http://www.geocities.jp/patentsearch2006/DB-table.html

6. 化合物関連データベース

In Silico Data
Miracles by the KY-methods

◇化合物・反応データベース

ﾃﾞｰﾀﾍﾞｰｽ名	提供元	収録年	収録件数	収録内容
化学物質・化学品				
REGISTRY	STN	1907-	84300000	CAﾌｧｲﾙに索引された化学物質、US,EP,CAの既存化学物質台帳収載物質
BEILSTEIN	STN/Dialog	1771-	9430000	有機化合物、有機金属化合物の物性・合成・反応情報
CASREACT	STN	1985-	1300000	CAの有機化学反応情報（収録：文献単位）
ODOOS	chem-station			有機合成反応ﾃﾞｰﾀﾍﾞｰｽ
C1DB Search System				C1反応ﾃﾞｰﾀﾍﾞｰｽ
CHEMCATS	STN		8870000	市販の化学薬品、その供給業者
CSCHEM	STN		187000	化学薬品、化学製品、供給業者、商品名
製品検索　物質検索	日本化学工業協会			製造輸入販売の化学製品の成分、用途、危険有害性分類、製品名、成分名、CAS番号で検索
日化辞Web	科学技術振興機構			化学物質辞書で文字列検索のほか構造検索もできる
NIMS物質材料ﾃﾞｰﾀﾍﾞｰｽ PolyInfo 高分子DB	物質・材料研究機構 (NIMS)			高分子、金属材料など各種のﾃﾞｰﾀﾍﾞｰｽ
SDBS	産業技術総合研究所			有機化合物のｽﾍﾟｸﾄﾙﾃﾞｰﾀﾍﾞｰｽ
chem.com				Bulk化学物質（構造検索もできる）、供給先
ChemNet				化学品Directory検索(化学品の売買)
ChemExper				化学品Directory検索(化学品の売買)
ChemFinder				化学物質・有機合成に関するﾃﾞｰﾀﾍﾞｰｽ
RAPRA	STN/Dialog	1972-		ﾌﾟﾗｽﾁｯｸ、ｺﾞﾑなど高分子ﾃﾞｰﾀﾍﾞｰｽ
rapra　polymerlibrary	rapra technology			ﾌﾟﾗｽﾁｯｸ、ｺﾞﾑなど高分子ﾃﾞｰﾀﾍﾞｰｽ(ｲﾝﾀｰﾈｯﾄ版)
KOMPASS USA	Dialog			米国の製造業、関連ｻｰﾋﾞｽ業の企業･製品
DIRECTORY OF CHEMICAL PRODUCTS	Dialog			世界の化学品・ﾎﾟﾘﾏｰﾒｰｶｰの製品
EnplaNet				樹脂、成形加工など
Plascom				ﾌﾟﾗｽﾁｯｸ製品情報
rubberstation	加藤事務所			ｺﾞﾑ製品情報
14906の化学商品	化学工業日報社			国内化学商品の用途・製造業者・生産量(冊子体・年刊)
内外化学品資料	ＣＭＣ			汎用樹脂、原料、有機中間体などの需給統計、価格など(冊子体・年刊)
Chemical Information Services				世界化学品年鑑、化学中間体化合物年鑑(固定料金制)

http://www.geocities.jp/patentsearch2006/DB-table.html

6. 化合物関連データベース

◇安全性データベース

Ariel WebInsight				世界の既存化学物質台帳(年間定額制)
日本ｹﾐｶﾙﾃﾞｰﾀﾍﾞｰｽ			130000	国内流通の化学品、安全、法規制ﾃﾞｰﾀﾍﾞｰｽ (年間定額制)
CHEMLIST	STN	1979-	207500	日本、米国、EC、ｶﾅﾀﾞ、韓国、ｵｰｽﾄﾗﾘｱの化学物質規制台帳情報
HSDB	STN		4500	化学物質の毒性、環境影響に関する文献情報
RTECS	STN/Dialog	1971-	162000	化学物質の毒性ﾃﾞｰﾀ、出典情報
CHEMSAFE			21400	可燃性物質(その混合物)の安全性関連物性ﾃﾞｰﾀ、文献
TOXLINE	STN/Dialog	1965-	2455000	MEDLINE,BIOSISなどから薬理作用、生理作用、毒性に関する文献情報
TOXLIT	STN	1965-	2494000	CAの薬理作用、生理作用、毒性文献情報
TOXNET	SIS			HSDB,TOXLINE 毒性関連ﾃﾞｰﾀﾍﾞｰｽ多数収録
危険物質ﾃﾞｰﾀﾊﾞﾝｸ	SIS			
OSHA	米国労働安全衛生局			安全・健康関係のﾃﾞｰﾀﾍﾞｰｽ
JETOC Index	化学物質安全情報ｾﾝﾀ			化学物質の安全性に関するﾃﾞｰﾀﾍﾞｰｽ
化学物質総合検索ｼｽﾃﾑ	製品評価技術基盤機構			化学物質の番号や名称等から、有害性、法規制、国際機関によるﾘｽｸ評価情報等を検索
PRTR制度対象物質DB	製品評価技術基盤機構			PRTR制度の対象物質を一覧表示
既存化学物質安全性点検ﾃﾞｰﾀ	製品評価技術基盤機構			
化学物質毒性DB	GINC(Global Information Network on Chemicals)			化学物質毒性試験報告
安衛法化学物質	中央労働災害防止協会			
化学物質の危険有害性	中央労働災害防止協会			
化学物質情報(kis-net)	神奈川県環境科学ｾﾝﾀｰ			
国際化学物質安全性ｶｰﾄﾞ(ICSC)	国立医薬品食品衛生研究所			
化学物質ﾌｧｸﾄｼｰﾄ	環境省			
PRTR法指定化学物質	環境省			

http://www.geocities.jp/patentsearch2006/DB-table.html

6. 化合物関連データベース

◇MSDS関連データベース

MSDS-OHS	STN	1984-	57000	米国のMSDS
MSDS-CCOHS	STN		141000	ｶﾅﾀﾞのMSDS(製造業者連絡先を含む)
MSDS on the internet				MSDSｻｲﾄのﾘﾝｸ集
日本試薬協会MSDS検索	日本試薬協会			国内試薬会社各社のMSDSを検索できる
ﾒﾙｸ･ｼﾞｬﾊﾟﾝ MSDS検索	ﾒﾙｸ･ｼﾞｬﾊﾟﾝ			

http://www.geocities.jp/patentsearch2006/DB-table.html

6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベース
◆連結型データベース
◆リレーショナルデータベース
◇SQL (Structured Query Lang

・階層型モデル

階層型モデルではデータの相互関係が階層的になり、且つデータ間の横のつながりが比較的弱いようなデータ構造を持つものに的したデータベースである。 このモデルのデータ間のつながりの形は基本的に木（ＴＲＥＥ）構造をなしている。 この木構造にも様々

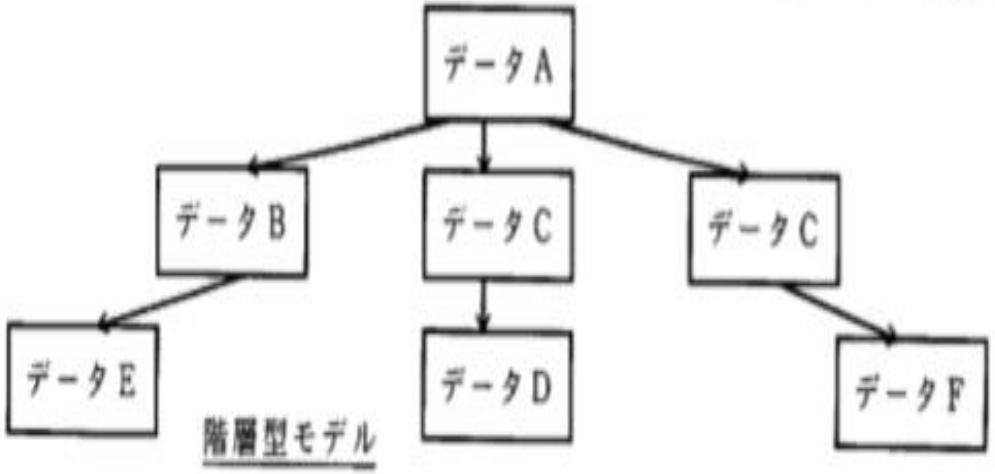


図 ．階層型モデルのデータ構造

6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベース
◆連結型データベース
◆リレーショナルデータベース
◇SQL (Structured Query Lang

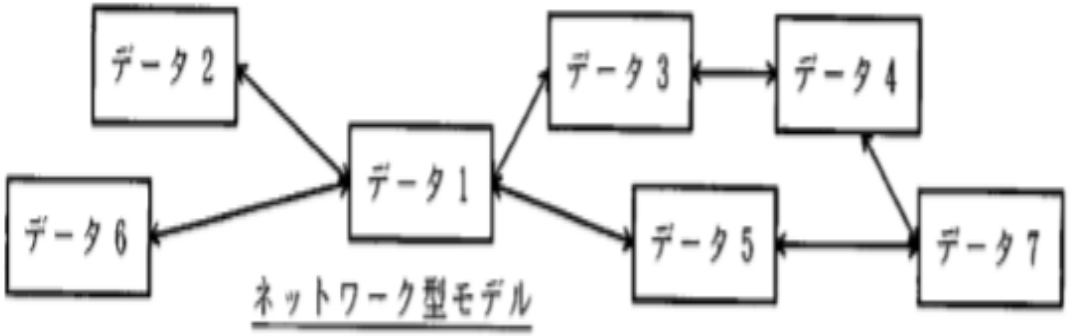


図　．ネットワーク型モデルのデータ構造

ネットワーク型モデルはより複雑な相互関係を持つデータを扱う時に利用されるモデルである。　個々のデータ間に高度な階層関係や相互関係が存在し、これらが互いに交錯しているようなデータを扱う時に利用される。

6. 化合物関連データベース

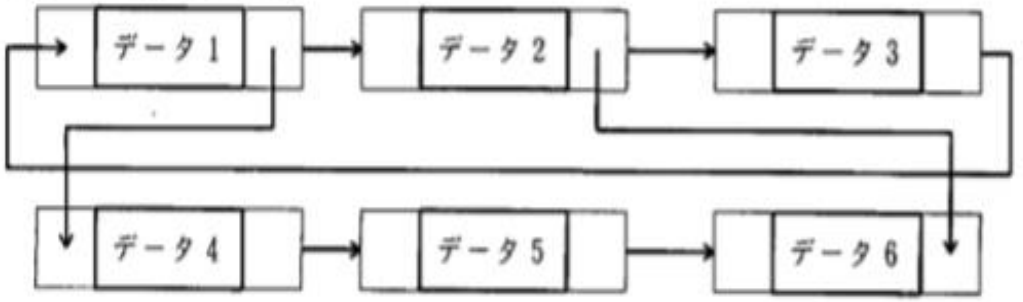
＊データベースの基本

◆階層型データベース
◆ネットワーク型データベー
◆連結型データベース
◆リレーショナルデータベー
◇SQL (Structured Query La

□　多重つなぎ編成によるデータベース構築

この多重つなぎ編成はネットワーク型モデルの一種である。　このアプローチではデータ間のネットワーク関係（数学的にはリスト：ＬＩＳＴと呼ばれ、データベース関連では鎖：ＣＨＡＩＮと呼ばれる）をそのままデータベース構造に反映させてデータベースを構築するものである。　従ってこのデータベースは高速化を狙うよりも、データ間の結合関係を重視したデータ検索を目的としている。

データは本来のデータ自身に対する情報と、そのデータと関連するデータへの結合情報部分とから構成される。



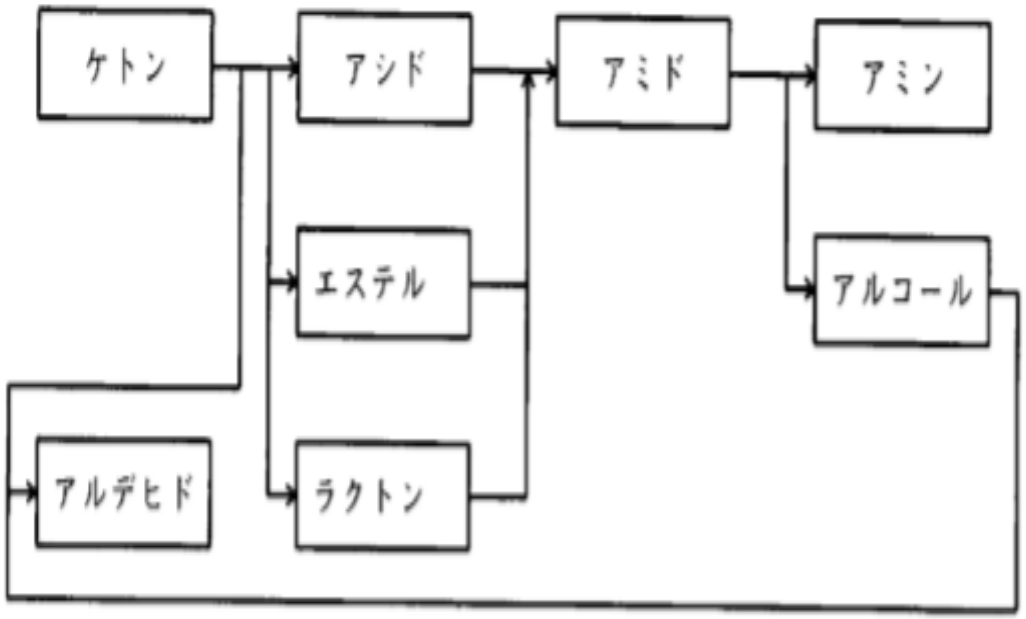
＊太線枠内は保存用データ、細線枠内が結合情報を持っている部分

6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベース
◆連結型データベース
◆リレーショナルデータベース
◇SQL (Structured Query Language)

例）ケトンを基点とした官能基群を考慮し、化合物データベースを構築する



6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベー
◆連結型データベース
◆リレーショナルデータベー
◇SQL (Structured Query Language)

$$R(X_1, X_2, \cdots\cdots X_n) = \{(x_1, x_2, \cdots\cdots x_n) \mid x_1 \in X_1 \wedge \text{命題} R(x_1, x_2, \cdots\cdots x_n) \text{が真}\} \subseteq X_1 \times X_2 \cdots X_n$$

X_i は関係の定義域（DOMAIN）、定義域の数を関係の度数（DEGREE）、
x_i は関係の要素、$(x_1, x_2, \cdots\cdots x_n)$ はn組（n－TUPLE）と呼ぶ。

第1マトリクス

関係名	定義域 1	2	3	4	5
実験者	実験者名	実験番号	実験日付	天気	実験結果
試薬	試薬名	購入日付	製造メーカ	残量	
化合物	化合物ＩＤ	製造者名	製造日付	機器分析	薬理データ

第2マトリクス

化合物	1 化合物ＩＤ	2 製造者名	3 製造日付	4 機器分析	5 薬理データ
化合物1	621A	KY	3.01.22	IR.NMR	11C
化合物2	372X	KK	3.02.15	IR.NMR	11C
化合物3	277C	TO	3.02.05	IR.MASS	12C

6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベース
◆連結型データベース
◆リレーショナルデータベース
◇SQL (Structured Query Language)

■その他のデータベース型
◆スプレッドシート型データベース
◆オブジェクト指向型データベース

表．ケミカルスプレッドシート例

化合物構造式	融点	沸点	薬理データ	LOGP
化合物 1	11.4	120.8	134	4.3
化合物 2	-5.8	186.4	155	2.7
化合物 3	123.3	256.1	176	3.3
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

6. 化合物関連データベース

＊データベースの基本

◆階層型データベース
◆ネットワーク型データベース
◆連結型データベース
◆リレーショナルデータベース
◇SQL (Structured Query Language)

■その他のデータベース型
◆スプレッドシート型データベース
◆オブジェクト指向型データベース

オブジェクトデータベースは、オブジェクト指向プログラミングで使うオブジェクトの形式で表現されるデータを格納するデータベースである。オブジェクト指向データベースともいう。

オブジェクト指向プログラミングにおいて、オブジェクトをデータベースに格納（永続化）する方法の一つである。オブジェクトデータベースは、オブジェクト指向プログラミング言語と密接に連携する。オブジェクトデータベースのデータベース管理システム (DBMS) を、オブジェクトデータベース管理システム (ODBMS; Object DBMS) 、あるいはオブジェクト指向データベース管理システム (OODBMS; Object Oriented DBMS) という。

https://ja.wikipedia.org/wiki/オブジェクトデータベース

In Silico Data
Miracles by the KY-methods

The Chem-Bio Informatics Society　情報計算化学生物学会

ご清聴ありがとうございました

株式会社　インシリコデータ
湯田　浩太郎