ケンブリッジ結晶構造データベースの使い方

(社)化学情報協会 千原秀昭

Hideaki CHIHARA: How to Use Cambridge Structural Database System

1. はじめに

Laue による X 線の発見から 82 年, Robertson がフーリエ級数法ではじめて有機化合物の結晶構造を決定してからでも 60 年, X 線結晶学はそれ自身長足の進歩を遂げるとともに, 自然科学のほとんどすべての分野の発展を支え, 要所要所で重要な発見をもたらしてきた。有機化合物と有機金属化合物に限っても, 膨大な結晶構造データの蓄積があって, X 線結晶学は物質の構造を知りたいときに使う手段としての使命のほか, 蓄積されたデータが分子設計, 反応設計などの基礎として新しい生命を見だしている。

本稿では, 有機化合物と有機金属化合物の結晶構造の集大成であるケンブリッジ結晶学データセンターのデータベース・システムの概略を紹介する。

歴史的背景: ケンブリッジ構造データベース (CSD) の集積の始まりは 1965 年にさかのぼる。この年ケンブリッジ結晶学データセンター (CCDC) が設立され, (a) 有機炭素化合物の結晶構造に関する数値データベースを構築し, (b) そのデータベースには遡及的に収集したデータを収容して, これを完全に評価し, (c) CSD に収録されたデータの検索, 分析, 表示ができるソフトを開発し, (d) それらのデータを利用した基礎研究を行うこととされた。CCDC は当初英国政府の財政的な援助を得て出発したが, いまでは完全な独立採算性をとっている。

データ収集が最大の目的であった最初のうちは結晶学の専門家によって, データを 3 個のファイルに分けて作成した。(1) FBIB (検索用は ABIB) は文献と化学的な文章情報 (2) FCON (検索用, ACON)

は二次元の結合表^{a)} (3) FDAT (検索用, ADAT) は三次元構造の数値情報の 3 ファイルで各ファイルは相互に CSD の参照コード (REFCODE)^{b)} で連結されている。センター内でデータ処理をするためにいろいろなプログラムが開発された。ABIB と ACON を検索するソフトを Version 1 として 1970 年代後半に, これを改良した Version 2 が 1980 年代はじめに公開された。これらはバッチ検索用のソフトで, Version 2 では VAX コンピュータ版がつくられた。その後間もなく, 3 ファイルを統合した ANEW ファイルも作られた。1988 年 1 月にはコマンド・モードで三次元検索が可能な Version 3, 1992 年には会話型グラフィックスで三次元検索ができ, 統計処理も可能な現行の Version 5 が公開された。ここでは Version 3 と 5 について紹介する。Version 4 は中間的なシステムなので, 説明は省略する。

CSD システムは構造データベースとその中の構造情報を検索, 表示, さらに計算や統計処理をするソフトからできている。データベースの部分は Version 3 でも 5 でもほとんど同じで, Version 5 では分子間の検索を可能にするために, matching number^{c)}が入っている点だけが異なっている。まず

-
- a) FCON ファイルには分子の構造式を表す結合表を納めてある。結合表は原子に番号をつけ, 例えば 2 番の原子は炭素で 3 番の原子 (例えば酸素) と 2 重結合で, 4 番の原子 (炭素) と単結合で結合しているということをマトリックスの形に表現したものである。参考文献 1) 参照。
 - b) 6 文字で化合物を表し, そのあとにその化合物の研究の歴史を示す 2 文字がつく。参照コードは CSD のどの Version でも共通である。
 - c) 結晶としての三次元構造と, その中の「分子」の構造とを対応させるためのリンク・キー。

データベースの内容の説明からはじめ、つぎにシステムとしてどんな機能があるかをVersion 3について述べ、Version 5でどんな機能強化が行われたかを紹介することにする。

2. ケンブリッジ構造データベース

データベースの内容は有機化合物、有機金属化合物、および有機リガンドをもつ金属錯体のX線、中性子線回折による研究結果で、完全に遡及する方針である(最も古いのは1930年のデータ)。年平均10,000件の率で追加されており、1993年11月の累計は116,413件の構造(化合物の数は104,106)を収容している。第1図はデータの増加率を示す。次第に大きな分子の構造データが増えていることがわかる。

各レコードには同定用の参照コード(REFCODE)がついている。収容されているデータの種別だけを並べると第1表のとおりである。ここで三次元データについて補足すると、結晶学的非対称単位には2個以上の分子やイオンが含まれていることも

あり、CSDではこれを「残基, residue」と呼ぶ。結晶学的な非対称単位に結合した原子は、対称操作で同位するものについても座標値を記録してある。

CSDを単なる結晶構造データの貯蔵庫でなく、化学的な研究に役立つものとするために障害となったのは、結晶構造を化学構造式と対応づけることであった。化学者がC=OあるいはC-Oとかくのが、三次元構造データでどのC $\cdots$ Oと対応するかは原子間距離や周囲の原子の配置から判定する。これにはかなりのソフト開発が必要であったが、その結果、matching numberを拡張結合表に組み込むことで解決した。これまでに87,568個の構造について完全なmatchingがとれており、10,181個についてはほぼ確かなmatchingが得られている。化学的な結合関係と結晶学的な結合関係が一致することを検証するので、matchingの作業はデータ自身の正確さとデータベースへの入力 of 正確さの厳しいチェックとなる。

第1図 a CSDの発行年とその年のデータ追加件数。b CSDの発行年とその年のデータ追加量(MB)。

第1表 CSDのデータの種類

文献データと 化学的データ	化合物名, 分子式, 著者名, 雑誌名, CODEN, 巻号頁年など, 説明語(中性子回折とか絶対配置の決定など), 実験法についてのコメント(普通でない方法の場合), 原報に記載の誤差, 構造の不整の有無, 化学クラス(数字で表現, ベンゼンのニトロ化合物なら15, 全部で86種に分類してある), 回折実験の精度, CSDにおけるデータ評価の状況。
二次元構造	分子の構造式と結合の種類を表す結合表 原子の性質: 原子につけた通し番号(n), 元素記号(el), それに結合する非水素原子の数(nca), 末端水素原子の数(nh), 正味電荷(ch), 立体構造を二次元で表示するための原子座標(x, yこれはVersion 5のみ)。 結合の性質: 結合で結ばれる原子対(i, j), 結合i-jの結合種(bt)。 結合種の種類: 単結合(bt=1), 二重(2), 三重(3), 四重(4), 芳香族(5), ポリマー(6), 非局在の二重(7), π 結合(9)。もしその結合が環の一部の場合は負号をつける。
三次元構造	結晶のデータ: 単位胞の大きさ, 空間群, 結合した結晶化学的単位の原子座標(x, y, z分数表示), 共有結合半径とその半径値から判別される結晶内の結合関係。

3. Version 3の機能^{1),2)}

あまりプログラムの詳細に立ち入らないで, Version 3の検索(探索^{d)})・表示機能をまとめることにする(Version 3の基本機能の解説は化学情報協会日本語版を準備中)。第2図がその概念図で, 上端の4種のデータを統合した探索用ファイルASERがある(Version 3では右端から1, 2番目の三次元立体図データは使えない)。これは上述のANEWファイルを再構成してビット・スクリーン形にして探索の高速化をはかったファイルである。探索プログラムはQUESTで英数字での探索ができる。探索結果のヒットはいろいろなファイルに収容するが, 特にFDATファイルには三次元データを収容し, あとでPLUTOプログラムで三次元グラフィックス表示をし, GSTATプログラムで計算や統計分析ができるようになっている。

QUESTプログラムはふつうの化学のデータベースの探索プログラムと同様な構造になっている。簡単な例を挙げると, SmithとJonesによって1986

年に発表されたステロイドの構造を捜すには, VAXコンピュータの場合, プロンプト(入力待ち記号)>のあとに次のように入力する。

```
>T1 *CLASS 51 (ステロイドはクラス51に分類されている)
>T2 *YEAR 1986
>T3 *AUTHOR Smith
>T4 *AUTHOR Jones
>QUES (T1.AND.T2).AND.(T3.OR.T4)
```

TはTESTと称する個別の質問単位で, TESTを組み合わせる最後の質問式(プロファイル) QUES (QUESTION)ができあがる。

このプログラムはFORTRAN77で書いてあるので, FORTRANになれた人は苦労せずにすぐ探索命令を書くことができる。たとえば測定密度が1.35以下のデータを捜すには

```
>T1 *DENM 0.0 - 1.35
>QUES T1
```

とすればよい。一つの残基の中の原子数を指定するときは, 元素記号がパラメーターとして入る:

```
>T1 *RESF Os 8 - 10 (Os原子が8-10個含まれるもの)
```

同様に, 文献や構造データのほとんどすべての項目についてTESTと, それを組み合わせ

d) 一般に search と retrieve の両方に対して「検索」という日本語が使われており, 区別するときに不都合を来している。本稿では区別のため, 前者は探索(質問に合致するデータを捜す), 後者は検索(見つけたデータを取り出す, 広義には探索を含む)とする。

第2図 CSD Version 3のフローチャート

QUESTを作って探索できる。このプログラムの特色は二次元構造(化学構造)の部分構造探索と類似探索にあるので、それを最後に紹介しよう。

二次元部分構造：いま簡単な例として

S-C-N=C

という部分構造(FRAGMENT)を含む構造を全部見たいでしょう。これは結合表(connectivity)を探索するので、次のように1組のパケットを指定する：

```
>T1 *CONN (結合表探索を指示する)
>>Q S-C-N=C FRAGMENT (コメント行)
>>AT1 S 1 (原子1番が硫黄で、少なくとも1個
            の非水素原子を結合している)
>>AT2 C 2
>>AT3 N 2
>>AT4 C 1
```

>>BO 1 2 1 (原子1番が原子2番と単結合している。最後の1は表1の結合種の番号)

>>BO 2 3 1

>>BO 3 4 2

>>END

類似探索：ある二つの化学構造が似ているかどうかは、両者の特徴を抽出し、両者に共通な特徴がどれだけあるかを調べればよい。共通な特徴が多数あるほど、この二つの構造は似ているわけである。CSDでは入力された構造を解析して特徴を抽出し、その1組の特徴を、データベースのなかに保存されている構造がもつ特徴と比較し、類似の程度を計算して、類似度の大きい順に100個の構造を検索する。入力する(質問の)構造はデータベースに存在しないものでもよい。

CSDでは434種の特徴をリストにして249から682までの番号をつけてある。たとえば、C-C結合で少なくとも4個の炭素原子がこれに結合している特徴を520番とする。ある分子が336番の特徴をもっていれば、336番目の桁を1と置き、その他の特徴がなければ、ほかの桁は0とする。このようにして作った434個の0または1の数字列をビット・スクリーンという。データベースに収録されているすべての構造について、構造のビット・スクリーンを与えてある。質問構造が入力されるとそれについて新たにビット・スクリーンを作り、これを保存してあるビット・スクリーンと比較する。いまビットが1であるスクリーンが質問構造について N_q 個、データベース中のある構造では N_i 個あるとする。両者に共通に1のビットが立っているスクリーンが N_c 個あったとすると、類似度は

$$T_{iq} = N_c / (N_q + N_i - N_c)$$

で表すことができる。これをタニモト係数^{e)}という。 T_{iq} が大きいかほどよく似ていることになる。第3図は類似度探索の一例である。実は部分構造探索はまずスクリーン探索によって、候補の構造を選び、つぎに詳しく結合表を比較している。これによって探索時間を大幅に短縮できる。

Version 3ではGSTATプログラムを別に利用すれば、三次元構造の検索や原子間距離の計算などができるが、この機能はVersion 5でメニュー方式に変わって、非常に便利になっている。GSTATを使えば、検索した一群の構造における $C(sp^2)$ -Cl結合の平均の長さなどを計算させることもできる。

4. Version 5の機能^{3),4)}

1992年に公開されたVersion 5では、QUESTとGSTAT, PLUTOプログラムがメニュー化されて、一体としてマウスで動くようになり、三次元構造検索がグラフィックス画面と英数字のコマンドとの組み合わせで手早く行えるようになった(QUEST3D)。構造の組立てはグラフィックス画面でメニューか

e) 類似の程度を表すパラメーターはいくつか提案されている。たとえば、
Willet, P; Similarity and Clustering in Chemical Information Systems. Research Studies Press-Wiley, Letchworth, U.K. 参照

質問構造

REFCODE=ZZZSE01

$T_{iq}=0.879$ (3位)

REFCODE=MORPIN

$T_{iq}=0.884$ (1位)

REFCODE=NALOXIC

$T_{iq}=0.729$ (55位)

第3図 類似探索の例。類似度 T_{iq} が大きいほど似た構造である。この例では質問構造はモルヒネ。

ら部品(原子, 原子団)を選ぶ方法でも, ある既存の構造から出発してそれを変更する方法でもできる。さらに原子間距離(分子内, 分子間), 結合角, ねじれ角の値の範囲を指定できる。環構造については環の中央(centroid)を定義して, その中央点から他の原子までの距離を範囲指定することができる。探索によってヒットした解答は1個ずつ画面に表示され(第4図), そのうち任意の構造を自分で名前をつけたファイルに保存して後でいろいろな解析に使うことができる。グラフィックス画面に表示した三次元構造図は拡大, 縮小, 回転させることができる。また表示中の構造について, 原子間距離や角度を計算させて画面下部に表示できる。

QUEST 3DプログラムにはGSTAT, PLUTOプログラムを包含したので, べつべつに起動する必要のあるVersion 3にくらべて非常に機動力に富むものになった。特に便利なのは, 非結合原子間の距離を指定した探索や, 検索された構造についての計算がその場でできる点で, これは分子内でも, 分子間でもよい。水素結合距離に対して威力を発

第4図 Version 5のグラフィックス表示画面

揮する。結晶内に2種以上の残基が含まれているような場合、たとえば、DNAの一部であるオリゴヌクレオチドと小さな薬品分子との相互作用の研究に応用できる。

5. 結 語

結晶構造決定が専門家の手を離れて、小さな分子ならば比較的簡単に誰でも構造を決定できるようになった事情を考慮して、自己のデータをCSDのASERファイルの形式で登録し、CSDのデータと同様に探索、検索ができるようになった。またCSDが結晶学者だけでなく、薬学、生化学の分野の研究者によって分子設計のための標準装備として使われるようになって、急速に普及している。このことを考慮して、同じく分子設計に広く使われているMDL社のMACCSシステムでも使えるフォーマットにASERファイルを変換するサービスも始まろうとしている。これと関連して、国際純正および応用化学連合(IUPAC)では分子構造の結

合表についての標準化の検討が進んでおり、いろいろなシステムが共通フォーマットのデータベースを利用できる方向に向かっている。CSDはブルックヘブン国立研究所のプロテイン・データバンクとも協力して、データの収録作業が無駄な重複とならないように、またデータのある程度共通にできるようにする調整が行われている。世界の主要な雑誌で行われているように、日本化学会の欧文誌(BCSJ)に掲載される結晶構造データを組織的にCSDに収録する方向での検討も進められている。

文 献

- 1) F. H. Allen, *et al*: J. Chem. Inf. Comp. Sci. **31**, 187 (1991).
- 2) CSD Version 3.4 QUEST Manual (1989).
- 3) O. Kennard: 包接と分子認識国際会議における講演, 京都, (1992).
- 4) CSD Version 5 QUEST 3D Manual (1992).