STRUKTURNÍ DATABÁZE ORGANICKÝCH A ANORGANICKÝCH STRUKTUR

J. Hašek

Institute of Macromolecular Chemistry AS CR, Heyrovského nám.2, 162 00, Prague 6

Keywords:
structural databases, organic compounds, organometalic compounds, proteins, structure analys

Abstract
The database of experimentally determined structures of organic and organometallic compounds containing more than 400 thousands of structures is a basic source of information on the detailed structure of molecules necessary in chemistry, pharmacology, structure biochemistry and technology. The paper reviews the scientific tools offered by CSD and gives information on availability of database and the related software in the Czech Republic.

Abstrakt
Strukturní databáze experimentálně stanovených organických a organometalických sloučenin, obsahující nyní více než 400 tisíc experimentálně stanovených struktur, je základním zdrojem informací o detailní struktuře molekul, které jsou nezbytné v chemii, farmakologii, strukturní biochemii, technologii. Článek podává přehled o možnostech, které současný systém Cambridgeské strukturní databáze (CSD) poskytuje a informuje o dostupnosti strukturní databáze a navazujícího sofwaru v České republice.

1. Úvod

Cambridgeská strukturní databáze je základním zdrojem informací o molekulární struktuře a mezimolekulárních interakcích experimentálně ověřených v pevné fázi. V současné době databáze obsahuje 401 tisíc sloučenin. Zhruba ve 230 tisíci případech jde o strukturu komplexů vytvářených organickými molekulami. Software dodávaný v rámci systému CSD umožňuje poměrně snadnou analýzu těchto mezimolekulárních interakcí a umožňuje tak například racionální přístup k návrhu nových dobře strukturně definovaných molekulárních systémů.

Většina struktur obsažených v databázi je stanovena rentgenovou difrakcí na monokrystalových vzorcích. Menší část (1230 struktur) je stanovena pomocí práškové difrakce nebo pomocí neutronové difrakce (1331 struktur). Je vhodné mít na paměti, že experimentálním výsledkem řešení struktury pomocí rtg difrakce je rozložení elektronové hustoty v molekule a tedy, že polohy atomů uváděné v databázi jsou středy jejich elektronových hustot. U neutronové difrakce jsou výsledkem získaným z experimentu polohy atomových jader. Tento fakt je významný zejména pokud nám jde o vodíkové atomy, kde u rtg difrakce vidíme elektronovou hustotu centrovanou v jiné poloze než je proton měřený neutronovou difrakcí.

Také je vhodné připomenout, že rtg difrakce jednoznačně řeší problém absolutní konfigurace a že databaze obsahuje 6291 jednoznačně stanovených opticky aktivních sloučenin. Z hlediska studia slabých mezimolekulárních interakcí je CSD atraktivním zdrojem nových poznatků ukazujících realitu. Většina objevovaných interakcí „nového typu“ byla zpravidla zřejmých již mnoho desítek let před jejich objevem prostým prohlédnutím experimentálně naměřených struktur obsažených v CSD.

2. Historie Cambridgeské strukturní database

Sběr dat do databáze organických a organometalických struktur stanovených metodami rtg difrakce byl iniciován Dr Olgou Kennardovou, vedoucí oddělení organické, anorganické a teoretické chemie Cambridgeské university, v roce 1964. Projekt byl zpočátku podpořen Britskými akademickými granty od “UK Office for Scientific and Technical Information”, dále “UK Science and Engineering Research Council”. V té době bylo několik konkurenčních projektů sbírajících obdobná data. Cambridgeské laboratoři se však podařilo získat mezinárodní podporu. Národní prostředky financování byly doplněny subvencemi tak zvaných Národních center CSD (“National Affiliated Centres”) zřízených v dalších čtyřech zemích (USA, Italie, Japonsko, Československo).

Začátkem sedmdesátých let začala mezinárodní distribuce prvních plně funkčních systémů pro vyhledávání a zpracování dat z této databáze a v roce 1980 už byla CSD lokalizována ve 30 zemích po celém světě. V této době o CSD systém projevily zájem i farmaceutické firmy a firmy produkující v oblasti chemie využívané v zemědělství a začaly projekt finančně podporovat. To umožnilo v roce 1998 přechod CSD na samostatnou neziskovou společnost (registered charity), která přijala jméno Cambridge Crystallographic Data Center (CCDC). Přestože středisko sídlí od roku 1992 v samostatných prostorách má organizačně velmi úzké spojení s Department of Chemistry Cambridgeské university a zajiš•uje zde část výuky. V současné době je databázový systém CSD distribuován do ca 60 zemí. CSD byla v letech 1965-1997 vedena dr. Kenardovou, v letech 1997-2002 Dr. Davidem Hartleyem a od roku 2002 Dr.Frankem Allenem. Část vývoje softwaru probíhá od roku 1998 v samostatné společnosti CCDC Software Limited. Dále CCDC nabízí pronájem vybraného softwaru, který byl vyvinut jinými společnostmi.

Stručný přehled historie

1964 O.Kennard – počátky sběru strukturních dat na Cambridgeské universitě
1974 Plně funkční systém CSD
- zřízení „Přidružených národních center CSD“ v USA, GB, Italii, Japonsku a Československu
- instalace CSD v ÚMCH AV, FZÚ AV ČR, MU Brno a STU Bratislava

1980 CSD lokalizována již ve 30 zemích světa
1997 D.Hartley ředitel

1998 Založena nezisková společnost CCDC (charitativní organizace)
2002 F.Allen ředitel
2007 CSD je nyní lokalizována v 60 zemích světa

3. Strukturní databáze organických a organometalických sloučenin v českých zemích

Československo bylo v roce 1974 mezi prvními pěti zakládajícími „Přidruženými národními centry“ (NAC - National Affiliated Centers), na kterých je činnost současné CCDC založena, zejména díky aktivitám Karla Humla z ÚMCH AV ČR. Od té doby byly lokální verse CSD vždy dostupné alespoň na třech pracovištích v Praze, v Brně a Bratislavě. V pozdějších letech se na zajištění provozu NAC podíleli Václav Langer, Bohdan Schneider, Jarmila Dušková a Jindřich Hašek. V letech 1993-2004 se o zajištění CSD v ČR významně zasloužili prostřednictvím svých grantů Jan Fábry (FZÚ AV ČR) a Jan Ondráček (ÚMG AV ČR).

Pro všeobecný přístup do databáze je nejdůležitější její sí•ová verse. Ta je již od roku 1974 dostupná každému nekomerčnímu zájemci v ČR přes počítačový server ve FZÚ AV ČR. O nepřetržitý provoz této sí•ové verse CSD ve FZÚ až do dnešní doby se zasloužili především Jaroslav Nadrchal, Alan Línek, Václav Petříček, Jan Fábry a Michal Dušek. V současné době je provoz sí•ové verse zajiš•ován Michalem Duškem (FZÚ AV ČR) a lokální instalace CSD jsou instalovány na deseti dalších pracovištích.

Cambridgeskou strukturní databázi tedy mohou v její sí•ové formě využívat všichni nekomerční zájemci v našich zemích nepřetržitě již od roku 1974. Do roku 2004 byl pronájem hrazen z prostředků státu a její využívání bylo bezplatné. V současné době nepovažují grantové agentury zajištění CSD pro celý stát za samostatný vědecký problém a vyžadují, aby si strukturní data zajistil každý řešitel vědeckého problému ze svých vlastních zdrojů.

Proto je od roku 2005 pronájem CSD hrazen z prostředků Krystalografické společnosti. Náklady s tím spojené jsou kompenzovány poplatky vyžadovanými od trvalých uživatelů CSD. Za občasné využití CSD po síti není žádný poplatek vyžadován, ale protože výše poplatků závisí na počtu platících uživatelů nemělo by být této možnosti nadměrně využíváno. V současné době je poplatek tvořen ze dvou částí. Výše první části závisí na odhadu počtu grantů využívajících CSD na daném pracovišti a druhá část je poplatek za každou samostatnou instalaci CSD na vlastním počítači.

Způsob úhrady poplatku za pronájem CSD v našich zemích
1974 – 1992 Akademie věd
1993 – 2004 Krystalografická společnost prostřednictvím podpory GA ČR
2005 – 2007 Krystalografická společnost z vlastních prostředků

Pronájem CSD je poskytován vždy na dobu jednoho roku. Obvykle na jaře každého roku je všem registrovaným uživatelům jednorázově obnoven software a celá databáze. V průběhu roku mají vlastníci licencí možnost třikrát doplnit svoji databázi čtvrtletními přírůstky nově zpracovaných struktur po síti přímo z CCDC v Cambridge.

Distribuce CSD v rámce národní licence je zajiš•ována prostřednictvím Národního centra CCDC „Affiliated center of CCDC“. Kontakt: Dr. Jindřich Hašek, ÚMCH AV ČR, Heyrovského nám.2, 162 06 Praha 6; tel: 296 809 390; fax: 296 809 410; email: hasek@imc.cas.cz.

Pokud chcete instalaci **CSD na svém vlastním počítači**, kontaktujte přímo Národní centreum CCDC. Pokud chcete využívat **CSD prostřednictvím počítačové sítě,** přečtěte si instrukce na

http://www-xray.fzu.cz/csd/csd.html a zkontaktujte Dr. Michala Duška z Fyzikálního ústavu AV ČR e-mail: dusek@fzu.cz

4. Přehled softwaru užívaného ve spojení s CSD

Většina softwaru spolupracujícího s Cambridgeskou strukturní databází je k disposici jak pro systémy Windows, tak pro UNIX. Pokud ne, je to uvedeno na konci popisu.

4.1. Programové vybavení CSD systému

ConQuest
Základní vyhledávací program v Cambridgeské databázi. Program je velice intuitivní a uživatelsky příjemný umožňující přístup k většině funkcí bez čtení manuálu. Nicméně pokud chcete plně využít možností stastistického zpracování a zobrazování, bez dodatečných rad a “neklikacího” programu Quest se možná neobejdete.

Mercury
Program je skutečně velice dobrý zejména pro analýzu a zobrazování mezimolekulárních kontaktů v krystalu. Umožňuje snadné a rychlé porovnávání velkého množství molekulárních struktur.

VISTA
Specializovaný tabulkový processor (obdoba známého Excellu) usnadňujícího práci, statistické vyhodnocení a tvorbu grafů z dat získaných ze strukturní databáze a nebo vypočtených na základě nalezenych struktur.

Mogul
Program pro zobrazování krystalových struktur, výpočet a statistické porovnání vybraných molekulárních descriptorů

IsoStar (server – UNIX only)
Program umožňující statistické zpracování, názorné zobrazování typických mezimolekulárních interakcí v krystalech. Několik tisíc funkčních skupin je již v databázi zpracováno a lze snadno vyhledávat a zobrazovat typické mezimolekulární interakce.

Quest
Předchůdce “klikacího” programu ConQuest pro vyhledávání podobných struktur nebo strukturních fragmentů v Cambridgeské databázi.

PREQUEST
Program umožňuje práci s vlastními dosud nepublikovanými daty tak, jako by již v databázi byla obsažena. Umožňuje kontrolu a transformaci vlastních strukturních dat do formátu CCDC, tj. vytvoří novou doplňkovou databázi *.aser, odpovídající 2D chemické diagramy, atd. z

dat ve formátech SHELX, CIF, SD nebo MOL2. Dostupné jen pro systém UNIX.

Rpluto
Grafický program na zobrazování molekul s mnoha vlastnostmi, které krystalograf ocení.

DASH
Program pro řešení krystalových struktur z práškových difraktogramů. Na vstupu vyžaduje správně oindexovaný difrakční záznam, základní buňku, znalost správné prostorové grupy a model molekuly s několika neznámými torsními úhly. 4.2. Programy usnadňující analýzu struktury proteinů

SuperStar
Program umožňující identifikaci oblastí na povrchu proteinu, které mají silnéí interakce s proteinem. Program využívá experimentálně získané statistické informace o četnosti typů interakcí z Cambridgeské strukturní databáze, z Proteinové strukturní databáze a nebo CSD data optimalizována pomocí Gaussianu. Dostupné jen pro systém UNIX.

ReliBase+
Databáze umožňuje vyhledávání proteinových struktur v PDB, analýzu, zobrazování mezimolekulárních interakcí, hydratace, prázdných prostor v komplexech makromolekulárních struktur obsažených v „Proteinové Strukturní Databázi (PDB). Systém má vlastní scriptový jazyk a obsahuje řadu předzpracovaných dat, které v PDB nejsou přímo obsažené. Dostupné jen pro systém UNIX.

Relibase
Relibase je neplacená varianta Relibase+ s omezenou funkčností. Dostupné jen pro systém UNIX.

GOLD
Gold (Genetic Optimisation for Ligand Docking) je program pro optimalizaci umístění flexibilního ligandu do vazebného místa proteinu. Scoring funkce jsou voleny tak, že dávají dobrou shodu (>70 %) s experimentálně stanovenými strukturami v PDB.

SILVER
Zobrazování, kontrola a detailní analýza interakcí mezi proteinem a ligandem. Vhodný zejména pro práci s programem GOLD

EnCIFer
Program usnadňuje přípravu CIF souborů potřebných pro publikaci a depozici krystalových struktur do CSD a do PDB. Deskriptory, které je třeba užívat pro organické, anorganická a ostatní struktury s malou základní buňkou jsou rozděleny do následujících slovníků:

- Core dictionary (coreCIF)
- Powder dictionary (pdCIF)
- Modulated and composite structures dictionary (msCIF)
- Electron density dictionary (rhoCIF)
- Descriptory používané pro makromolekulární struktury jsou popsané ve slovnících:
- Macromolecular dictionary (mmCIF)
- Image dictionary (imgCIF)
- Symmetry dictionary (symCIF)

Vysvětlení formátu CIF a popis významu deskriptorů lze nalézt na adrese
http://www.ccdc.cam.ac.uk/products/csd/faqs/#deposition

4.3. Doporučený způsob odkazů na využití CSD systému

CSD system F.H.Allen and W.D.S.Motherwell, *Acta Crystallogr.,* B58, 407-422, 2002.
ConQuest, Mercury Bruno I.J. et al. *Acta Cryst.* (2002), B58, 389-397.
IsoStar Bruno I.J. et al. *J.Comp.Aided Mol.Design* (1997),11-6, 389-397.
SuperStar Verdonk M.I. et al. *J.Mol.Biol* (1999) 289, 1093-1108.
Mogul Bruno I.J. et al. *J.Chem.Information and Computer Science.* (2005), 44-6, 2133-2144.
Life Sciences R. Taylor, *Acta Crystallogr.,* **D58,** 879-888, 2002.

5. Závěr

Cambridgeská strukturní databáze obsahující vice než 400 tisíc experimentálně stanovených krystalových struktur organických sloučenin a jejich komplexů a je každoročně doplňována 30 tisíci nově stanovenými strukturami. Snadná manipulace a intuitivní ovládání většiny programů umožňuje využití databáze nejen ve vědecké práci, ale též při výuce na vysokých a středních školách. Možnost snadného zobrazení reálné prostorové struktury molekul a možnost prohlížení jejich interakcí s okolními molekulami je nejen vysoce pozitivní z pedagogického hlediska, ale činí výuku chemie pro studenty názornější a atraktivnější.

Poděkování: *Projekt je částečně podporován z prostředků GA AV ČR IAA500500701.*

6. Literatura

Přestože je ovládání database intuitivní a ve velké většině "klikací", doporučujeme zájemcům prostudování manuálů volně dostupných na WWW stránkách. Manuály lze zdarma stáhnout z adresy
http://www.ccdc.cam.ac.uk/support/documentation/#csds

V následujícím seznamu literatury jsou další práce doporučené k přečtení setříděné tématicky podle jmen uživatelských programů.

Systém CSD

Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry.
F.H.Allen and W.D.S.Motherwell, *Acta Crystallogr.*, **B58**, 407-422, 2002.

The Cambridge Structural Database: a quarter of a million crystal structures and rising. F. H. Allen, *Acta Crystallogr.*, **B58**, 380-388, 2002.

Applications of the Cambridge Structural Database to molecular inorganic chemistry. A. G. Orpen, *Acta Crystallogr.*, **B58**, 398-406, 2002.

Life Science applications of the Cambridge Structural Database. R. Taylor, *Acta Crystallogr.*, **D58**, 879-888, 2002.

Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph. J. van de Streek, *Acta Cryst.*, **B62**, 567-579, 2006. [DOI: 10.1107/S0108768106019677]

ConQuest

New software for searching the Cambridge Structural Database and visualising crystal structures. I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallogr.,* B58, 389-397, 2002.

Mercury

Rules governing the crystal packing of mono- and di-alcohols R. Taylor and C. F. Macrae, *Acta Crystallogr.*, **B57**, 815-827, 2001.

VISTA

CCDC (1994). Vista - A Program for the Analysis and Display of Data Retrieved from the CSD. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

PREQUEST

PreQuest - A program for the validation of crystal structure and chemical information for entry to the CSD. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

ISOSTAR

Isostar: A library of information about non-bonded interactions. I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, *J. Comput.-Aided Mol. Des.*, **11**, 525-537, 1997.

The Protein Data Bank . H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 28, 235-242, 2000.

MOGUL

Retrieval of Crystallographically-Derived Molecular Geometry Information. I. J. Bruno, J. C. Cole, M. Kessler, Jie Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris and A. G. Orpen, *J. Chem. Inf. Comput. Sci.,* 44(6), 2133-2144, 2004.

SUPERSTAR

SuperStar: A knowledge-based approach for identifying interaction sites in proteins M. L. Verdonk, J. C. Cole and R. Taylor, *J. Mol. Biol.,* 289, 1093-1108, 1999.

SuperStar: Improved knowledge-based interaction fields for protein binding sites. M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet and P. Willett, *J. Mol. Biol.,* 307, 841-859, 2001.

SuperStar: Comparison of CSD and PDB-base interaction fields as a basis for the prediction of protein-ligand interactions. D. R. Boer, J. Kroon, J. C. Cole, B. Smith and M. L. Verdonk, *J. Mol. Biol.,* 312, 275-287, 2001.

Simple knowledge-based descriptors to predict protein-ligand interactions. Methodology and validation. J. W. M. Nissink, M. L. Verdonk and G. Klebe, *J. Comput.-Aided Mol. Des.,* 14, 787-803, 2000.

GOLD

Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. G. Jones, P. Willett and R. C. Glen, *J. Mol. Biol.,* 245, 43-53, 1995.

Development and Validation of a Genetic Algorithm for Flexible Docking. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 267, 727-748, 1997.

A new test set for validating predictions of protein-ligand interaction. J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins,* 49(4), 457-471, 2002.

Improved Protein-Ligand Docking Using GOLD. M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins,* 52, 609-623, 2003.

RELIBASE

Relibase - Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions. M. Hendlich, A. Bergner, J. Gunther, G. Klebe, *J. Mol. Biol.*, 326, 607-620, 2003.

Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase. J. Günther, A. Bergner, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 326, 621-636, 2003.

Use of Relibase for retrieving complex 3D interaction patterns including crystallographic packing effects. A. Bergner, J. Günther, M. Hendlich, G. Klebe and M. Verdonk, *Biopolymers (Nucleic Acid Sci.),* 61, 99-110, 2002.

DASH

Routine determination of molecular crystal structures from powder diffraction data. W. I. F. David, K. Shankland, N. Shankland, *Chem. Commun.,* 931-932, 1998.

R. E. Dinnebier, P. Sieger, H. Nar, K. Shankland, W. I. F. David, *J. Pharm. Sci.,* 89, 1465-1479, 2000.

H. Nowell, J. P. Atfield, J. C. Cole, P. J. Cox, K. Shankland, S. J. Maginn, W. D. S. Motherwell. *New J. Chem.*, 469-472, 2002.

A. Boultif, D. Louer, *J. Appl. Cryst.*, 24, 987-993, 1991.

CSD Symmetry

CSDSymmetry: the definitive database of point group and space group symmetry relationships in small- molecule crystal structures. J.W.Yao, J.C.Cole, E.Pidcock, F.H.Allen, J.A.K.Howard, W.D.S.Motherwell, Acta Cryst., B58, 640-646, 2002.

A Database Survey of Molecular and Crystallographic Symmetry. E. Pidcock, W. D. S. Motherwell, J. C. Cole, Acta Cryst., B59, 634-640, 2003.

ENCIFER

enCIFer: A program for viewing, editing and visualising CIFs. F. H. Allen, O. Johnson, G. P. Shields, B. R. Smith, M. Towler, J. Applied Cryst., 37, 331-334, 2004.