Chem. Listy *105*, 467–475 (2011) Referát

STRUKTURNÍ DATABÁZE ORGANICKÝCH A ORGANOMETALICKÝCH SLOUČENIN

JINDŘICH HAŠEK

Ústav makromolekulární chemie AV ČR, Heyrovského nám. 2, 162 06 Praha 6
hasek@imc.cas.cz

Klíčová slova: strukturní databáze, organické sloučeniny, organometalické sloučeniny, sydnony, crown-ethery

Došlo 1.12.09, přepracováno 5.8.10, přijato 2.12.10.

Obsah

1. Úvod
2. Dostupné databáze krystalových a molekulárních struktur
   2.1. Cambridgeská strukturní databáze
   2.2. Data uložená v CSD
   2.3. Přesnost a spolehlivost struktur obsažených v CSD
   2.4. Vodíkové atomy v CSD
   2.5. Další strukturní databáze
3. Software pracující s daty uloženými v databázi CSD
4. Programy využívající znalosti získané analýzou strukturních databází
5. Příklady aplikací Cambridgeské strukturní databáze
   5.1. Vlastnosti fotochromních sloučenin obsahujících sydnonový kruh – indukční a mezomerní efekt
   5.2. Interakce polyethylenoxidu s proteiny
6. Dostupnost CSD v ČR
7. Školní verze Cambridgeské strukturní databáze

1. Úvod

Strukturní databáze organických a organometalických struktur „Cambridge Structural Database (CSD)“ byla založena kolem roku 1965 z iniciativy Dr. Olgy Kennardové na katedře organické, anorganické a teoretické chemie na Cambridgeské univerzitě. Na rozdíl od několika podobných strukturních databází, které se v této době objevily a později zanikly, CSD byla v 70. letech přetransformována do mezinárodního projektu členských států, přispívajících každoročně na provoz CSD. Československo reprezentované Československou akademií věd se stalo jedním z prvních 5 zakládajících členů tohoto sdružení a celá nezisková vědecká komunita v České a Slovenské republice má od této doby k CSD přístup.

V roce 1980 byla CSD distribuována již do více než 30 členských států. V 90. letech minulého století byla CSD přetransformována na samostatnou neziskovou společnost nazvanou „Cambridge Crystallographic Data Centre (CCDC)” s více než 50 zaměstnanci sídlící v samostatné budově a pouze volně přidruženou ke Cambridgeské universitě. Po odchodu O. Kennardové vedl CSD v letech 1997–2002 Dr. David Hartley, v letech 2002–2008 Dr. Frank Allen a v současné době Dr. Colin Groom.

V současné době jsou CSD a příslušný software využívány v sedmdesáti zemích světa. Na vývoji CSD systému v roce 2007 spolupracovalo 55 pracovišť. Licence zakoupilo více než 1000 akademických pracovišť ze 69 zemí světa a více než 200 komerčních společností z 15 zemí. Mezi země s největším počtem instalací CSD patří USA (243), Japonsko (105), Francie (99), Německo (69), Ruská federace (59), Polsko (55), Španělsko (51), Itálie (40), Velká Britanie (37), Kanada (35).

2. Dostupné databáze krystalových a molekulárních struktur

2.1. Cambridgeská strukturní databáze

Strukturní databáze organických a organometalických struktur „Cambridge Structural Database (CSD)“ je distribuována střediskem „Cambridge Crystallographic Data Centre (CCDC)” a obsahuje všechny publikované struktury organických a organometalických látek určené experimentálně metodami difrakce rtg záření nebo difrakce neutronů. Deponované struktury musí vyhovovat kriteriím CCDC na přesnost a spolehlivost výsledků. V současné době obsahuje databáze CSD půl milionu organických a organometalických sloučenin. CCDC software umožňuje rychlé vyhledávání struktur, výpočet strukturních parametrů a statistickou analýzu zjištěných strukturních rozdílů. Do databáze jsou ukládána všechna strukturní data získaná difrakcí rtg záření a neutronů na sloučeninách, které obsahují v molekulární struktuře alespoň jeden uhlíkový atom, které nepatří do žádné z kategorií (strukturních databází) vyjmenovaných v následujícím odstavci a které splňují minimální požadavky kladené CCDC na přesnost a spolehlivost zjištěných výsledků.

2.2. Data uložená v CSD

Primární data uložená v CSD obsahují střední polohy atomů v asymetrické části základní buňky uložené v podobě frakčních souřadnic, údaje o teplotních kmitech všech atomů ve formě teplotních elipsoidů a standardní odchylky všech těchto experimentálně získaných dat. Dále jsou zde nezbytné informace o základní buňce a symetrii krystalu, chemické názvy sloučeniny, strukturní vzorec, odkazy na literaturu a fyzikální charakteristiky uvedené autorem ve zdrojové publikaci.

Data uložená v CSD nejsou určena k přímému čtení. To by bylo příliš obtížné. Intuitivně snadno ovládatelné programy umožňují uživateli rychlé a spolehlivé vyhledání všech struktur obsahujících uživatelem nakreslený molekulární fragment. To umožňuje získat množinu cílových struktur podle jeho požadavků on-line a formátovat je přímo do numerických nebo grafických výstupů. Lze také generovat vybraná data v nejrůznějších výstupních formátech pro zpracování výsledků v jiných počítačových programech (univerzální CIF, XML, PDB s orthonormálními souřadnicemi, XLS pro tabulkové procesory, atd.). Důvodů pro použití jiných programů však není mnoho, protože CSD software poskytne téměř vše požadované snadněji a rychleji. Jednotlivé struktury jsou v CSD systému jednoznačně identifikovány šestipísmenkovým kódem (tzv. CSD kódem) usnadňujícím manipulaci. Pokud uživatel zjistí, že se struktura s určitým CSD kódem nechová podle jeho představ, může pouhým kliknutím na příslušný CSD kód zobrazit příslušný strukturní vzorec, chemický název, fyzikální charakteristiky a analyzovat příčiny. Systém tedy umožňuje bezproblémové a systematické zpracování i mnoha tisíc strukturních charakteristik najednou.

Práce s CSD je do značné míry intuitivní a automatická a nevyžaduje speciální znalosti v oboru. Například zobrazení orientace molekuly uvnitř krystalu umožňující natočení krystalu tak, abychom se dívali podél vybrané polární meziatomové vazby v molekule, zvládne v tomto softwaru i běžný student základní školy bez znalostí krystalografie. Pokud uživatel najde tutéž strukturu měřenou při různých teplotách, může získat představu o fázových přechodech na atomární a molekulární úrovni nebo porovnat makroskopicky a mikroskopicky získané tenzory teplotní roztažnosti. Technicky zaměření studenti snadno zvládnou analýzu ukazující, které meziatomové vazby a nebo mezimolekulární interakce jsou odpovědné za to, že tepelná roztažnost zkoumaného materiálu je v určitém směru významně větší (anizotropie tepelné roztažnosti).

2.3. Přesnost a spolehlivost struktur obsažených v CSD

Přesnost molekulární struktury získané difrakčními metodami je zpravidla velmi vysoká. Očekávané chyby ve vzdálenostech jsou v tisícinách Å (tj. lepší než pikometr) a přesnost stanovení vazebných úhlů je v setinách stupně. Na úrovni této přesnosti můžeme spolehlivě pozorovat vlivy, které ve struktuře molekuly vyvolávají slabé mezimolekulární interakce, intra- i extramolekulární přenosy náboje, mezimolekulární interakce, tepelný pohyb molekul, atd. Ve strukturách můžeme pozorovat teplotní pohyb atomů, různou konformaci molekul v měřeném vzorku, různé mezimolekulární interakce, disorder, atd.

Přesnost vlastní metody rtg strukturní analýzy je mnohem vyšší než jsou skutečné rozdíly v geometrii jednotlivých molekul v reálném vzorku. Rtg strukturní analýza tedy popisuje průměrné rozložení elektronové hustoty, průměrné polohy atomů a reálnou strukturu materiálů včetně teplotního pohybu. Má to výhodu, že tyto efekty reálné struktury můžeme experimentálně měřit. Ovšem, pokud s těmito efekty nepočítáme, projeví se to při interpretaci jako rozptyl dat. Jinými slovy, přesnost dat v CSD je dána zejména kvalitou připraveného vzorku.

RTG difrakce dává dynamickou strukturu reálného molekulárního systému zahrnující teplotní pohyb atomů, různou konformaci molekul v měřeném vzorku, různé mezimolekulární interakce, disorder, atd. Abstraktní metody sledující modelové systémy, jako jsou molekulární modelování a kvantové výpočty molekul v homogenním dielektriku, dávají prakticky vždy geometrii lišící se od struktury v CSD o více než připouští směrodatné odchylky experimentu. To nám dává jedinečnou příležitost vysvětlovat rozdíly mezi teorií a skutečnou molekulární strukturou, tedy kvantifikovat vliv prostředí, ve kterém se molekula nachází, a vliv mezimolekulárních interakcí, ke kterým v reálném prostředí vždy dochází. Mezimolekulární interakce jsou v krystalovém prostředí zpravidla jednotné, a proto bývá tato analýza poměrně přímočará a nenáročná.

CSD poskytuje možnost sledovat závislost změn molekulární struktury na vlastnostech substituentů a na jejích interakcích se sousedními molekulami a tak poskytuje možnost ověřovat platnost teoreticky navržených stavů systému nebo reakčních mechanismů. Například je-li navržený přesun náboje nebo navržený elektrofilní nebo nukleofilní efekt patřičně reflektován v reálné struktuře. CSD obsahuje tisíce již zpracovaných závislostí pozorovaných změn geometrie vlivem mezimolekulárních interakcí (lze prohlížet pomocí programu ISOSTAR). Uživatel si může snadno vypracovat svoje vlastní analýzy, pokud v systému chybí (viz příklad změn geometrie sydnonového kruhu vlivem různých substituentů uvedený v kapitole 5.1).

Statistické zpracování výsledků získaných z CSD umožňuje zobecnit některé pozorované jevy i na materiály a skupenství, pro které nemáme srovnatelně spolehlivou a přesnou metodu. Zapomeňte na pověry o tajemném krystalovém poli. Programem MERCURY lze poměrně snadno zobrazit a analyzovat všechny významné mezimolekulární interakce zkoumaného fragmentu u velkého počtu struktur. Lze očekávat, že obdobné typy interakcí se budou kompetitivně projevovat i v odlišném prostředí. Příklad takového zobecnění pozorovaných interakcí na polymerní materiály a biologické makromolekuly je uveden v kapitole 5.2).

Lze také pozorovat posuny elektronové hustoty vůči poloze jader, například posun vazebných elektronů v silných vodíkových vazbách, atd. Porovnáním reálné struktury získané z CSD s kvantovými výpočty, které obvykle reprezentují stav molekuly v homogenním prostředí a v energetickém minimu, můžeme získat velice cenné informace o příčinách pozorovaných rozdílů mezi teoretickou a reálnou strukturou.

Statistické vyhodnocení trendů založené na větším počtu podobných struktur dává vždy přesvědčivější a přesnější obraz reality, umožňuje lépe sledovat korelace mezi sledovanými parametry a eliminovat vliv dalších rušivých faktorů na sledované vlastnosti a získat spolehlivější analýzu obecných zákonitostí.

2.4. Vodíkové atomy

Současná rtg strukturní analýza umožňuje spolehlivé stanovení elektronové hustoty každého elektronu, a tedy snížení elektronové hustoty v místě očekávaného „vodíku“ může znamenat významný přesun náboje a nebo smíšený (multikonformační) stav v daném místě. U struktur staršího data je však třeba počítat s tím, že při použití tehdejších postupů nebyly polohy vodíkových atomů tak spolehlivě určené jako dnes. Pokud je zajímavá analýza poloh vodíkových atomů, je v tomto smyslu vhodné buď data před rokem 1990 vyloučit nebo se obrátit na zkušeného krystalografa, který je ve většině případů schopen na základě dat deponovaných v CSD zobrazit experimentální mapu elektronové hustoty a tak případné nejasnosti odůvodnit.

Polohy atomů uváděné v CSD jsou polohami středů elektronové hustoty jednotlivých atomů. Projevuje se to zejména u vodíkových atomů, u nichž je střed elektronové hustoty jediného elektronu vysunut cca o 0,1–0,2 Å směrem k centrálnímu atomu, nebo naopak u silných vodíkových vazeb můžeme pozorovat výskyt elektronu až ve vzdálenosti 1,3 Å od obou nevodíkových atomů. U molekulárního modelování, kde jsou udávány polohy atomových jader, je tedy třeba vzít v úvahu, že elektron vodíkového atomu se účastní vazby, a proto je střed jeho elektronové hustoty uvedený v CSD posunut zpravidla o 10 % k centrálnímu atomu. Pokud uživatele zajímají přesné experimentální polohy protonů, musí ve svému požadavku na CSD zatrhnout, že chce pouze výsledky získané pomocí difrakce pomalých neutronů. Neutrony interagují pouze s magnetickým momentem atomových jader, a proto je u této metody stanovení experimentální polohy osamoceného protonu velice přesné a spolehlivé.

2.5. Další strukturní databáze

Strukturní data, která nejsou obsažena v Cambridgeské strukturní databázi, lze nalézt v následujících strukturních databázích:

- „Protein structure database“ (PDB)[5]. Zde jsou shromažďována data o strukturách proteinů, polypeptidů a polysacharidů s více než 24 residui (monomery). „Proteinovou strukturní databázi“ lze využívat bez licence například na adresách http://www.rcsb.org/pdb/ nebo http://www.ebi.ac.uk/pdbe-site/PDBeSite/
- „Nucleotide structure database“ (NDB). Strukturní data týkající se oligonukleotidů jsou shromažďována v „Databázi nukleových kyselin“. Využití zdarma je možné na adrese http://ndbserver.rutgers.edu/
- Anorganické struktury jsou dostupné v „Databázi anorganických krystalových struktur“. Využití této databáze podléhá licenci, kterou lze získat na adrese: http://www.fiz-karlsruhe.de/icsd_content.html
- Databáze struktur kovů a slitin. Strukturní data o kovech a slitinách jsou shromažďována v databázi „CRYSTMET®“ na adrese: http://www.tothcanada.com/databases.htm
- Databáze klasických organických polymerů (PolyBase)[6]. S pojmem polymerní řetězec se v CSD dále setkáváme obvykle tam, kde jsou v krystalu tvořeny nekonečné řetězce propojením organických molekul prostřednictvím atomů kovů. Pod kategorií „polymer“ najdete pouze 6 struktur klasických polymerů. Důvodem, proč strukturní data klasických polymerů v CSD nejsou obsažena, je to, že jejich kvalita a zpracování zpravidla neodpovídá normě požadované administrací CSD. Krystalická fáze polymerních materiálů potřebná ke strukturní analýze je u polymerů zpravidla špatně uspořádaná a metastabilní. Proto nejsou difrakční data struktur klasických polymerů z hlediska kvality klasického zpracování rtg strukturní analýzou dobrá. Difrakční obraz mívá u polymerů z důvodů jejich enormní flexibility a polydisperzity zpravidla jen malý počet velmi difuzních reflexí. Proto jsou upřesněné struktury zpravidla globálně správné, ale detaily struktury nebývají na první pohled přesné, pokud nejsou upřesněné jako teoretický model. Přístup k polymerním strukturám deponovaným v PolyBasi[6] lze získat na adrese hasek@imc.cas.cz.

3. Software pracující s daty uloženými v databázi CSD

Pro práci se strukturními daty deponovanými v CSD byl vyvinut velmi sofistikovaný a rychle pracující vyhledávač struktur (v současné době zvaný CONQUEST), bez něhož je práce s CSD prakticky vyloučena. Tento software je vylepšen programy pro doplnění databáze svými vlastními dosud nepublikovanými daty, pro prohlížení a prostorové manipulace a též pro základní statistické vyhodnocování korelací mezi strukturou a vlastnostmi materiálů.

CONQUEST – umožňuje vyhledávat struktury nejen podle chemického složení, sumárního vzorce, ale i podle výrazů použitých v textu. Dále poskytuje vyhledávání podle strukturního vzorce a uspořádání atomů či molekul v prostoru.

PREQUEST – umožňuje uživateli přidat do CSD databáze svoje vlastní (dosud nepublikované) struktury a provádět statistické analýzy s kompletní databází.

enCIFer – program pro editování a kontrolu CIF souboru (Crystallografic Information File). Tento soubor má formálně jednoznačně a velice přísně definovaný obsah a slouží u některých časopisů přímo pro zasílání článku k publikaci.

MERCURY – velmi praktický a snadno ovladatelný program pro zobrazování a grafickou analýzu stavby krystalu a pro analýzu molekulární struktury. Umožňuje vyhledávat podobné motivy uspořádání molekul v krystalu, analýzu sítí vodíkových vazeb v pevné fázi atd. http://www.ccdc.cam.ac.uk/support/documentation/mercury/2_2/portable/appc_tutorials.6.1.html

VISTA – program pro statistickou analýzu dat získaných z CSD. Specializovaný tabulkový procesor

s možností vhodné úpravy a analýzy získaných dat, tisk grafů, histogramů, atd. Data lze také exportovat do xls formátu a pracovat v jiných tabulkových a grafických programech.

Některé programy mají svoji částečně blokovanou volnou verzi fungující bez licence. Najdete je na www stránkách v části „free services“.

4. Programy využívající znalosti získané analýzou strukturních databází

CCDC distribuuje též řadu programů, které využívají poznatky získané analýzou strukturních dat deponovaných v CSD. Jde zejména o poměrně obsáhlé knihovny experimentálně zjištěných korelací mezi strukturou a vlastnostmi. Dále poskytuje modelovací programy, ve kterých jsou parametry empirického potenciálu (slangově „forcefieldu“) optimalizovány tak, aby výsledky byly v co nejlepší shodě s těmito experimentálně zjištěnými korelacemi. Jsou zde dále programy pracující s proteinovou strukturní databází a programy usnadňující práci s práškovými difrakčními záznamy.

MOGUL – Program, který zobrazuje statistiku požadovaných vazebných délek, úhlů nebo torzních úhlů pro zadaný strukturní fragment pro všechny struktury, ve kterých byl tento fragment nalezen. Struktury odpovídající jednotlivým bodům lze jediným kliknutím zobrazit a tak je možno snadno vyloučit nevhodné případy.

ISOSTAR – Objemná databáze statistických analýz mezimolekulárních interakcí pozorovaných v experimentálně stanovených molekulárních strukturách pomocí difrakčních metod. Systém v současné době obsahuje 22 161 grafických analýz závislostí strukturních parametrů získaných analýzou CSD, 7459 závislostí získaných z PDB a 1550 teoretických studií profilů potenciálové hyperplochy v okolí lokálních minim. Uživatel tak má k dispozici statistický popis všech experimentálně popsaných interakcí mezi 350 centrálními funkčními skupinami a 45 vybranými interagujícími funkčními skupinami. Uživatel též může vytvářet svoje vlastní analýzy, zjišťovat korelace a zobrazovat histogramy a mapy hustoty výskytu konfigurací požadovaných funkčních skupin s vybranými reagenty.

SUPERSTAR – Program pro identifikaci interakčních a katalytických míst ve strukturách proteinů. Trojrozměrné mapy pravděpodobnosti výskytu odvozené z experimentálních rtg struktur ukazují vhodná místa pro interakce vybraných ligandů s proteiny zvolené třídy (vazebná místa proteinů).

RELIBASE – Program usnadňující analýzu struktury proteinů. Obsahuje databázi mezimolekulárních interakcí vytvořenou z dat obsažených v "Proteinové strukturní databázi" obsahující v současné době téměř 50 tisíc makromolekulárních struktur stanovených difrakcí rtg záření. RELIBASE obsahuje přehledy vazebných míst pro ligandy (substráty, inhibitory), které byly experimentálně nalezeny rtg difrakcí.

HERMES – Program pro grafické znázornění a analýzu interakcí mezi proteiny a ligandy (program Merkury pro proteiny nelze použít). Je vhodný zejména pro práci se systémy SuperStar, Relibase, GOLD, Mogul a IsoStar.

GOLD (Genetic optimization for flexible ligand docking) – Program nalezení optimálního umístění ligandu v molekule proteinu na základě pseudopotenciálů nastavených tak, aby vypočítané modely souhlasily co nejlépe s experimentálně stanovenými strukturami v CSD. Formálně jsou použity atomové a vazebné typy známé z programu SYBYL, ale empirické potenciály (force fields) a geometrická omezení (constraints) jsou odlišná. Program je tedy optimalizován tak, aby dával střední geometrii obdobných fragmentů pozorovaných v Cambridgeské a Proteinové strukturní databázi[5]. Program vyžaduje kontrolu strukturních typů a pečlivé doplnění vodíkových atomů, ale počáteční poloha, orientace ligandů ani konformace bočních řetězců nejsou podstatné. Program používá „genetický algoritmus“ pro hledání optimální cesty pro vložení ligandu do vazebného místa v proteinu. Má optimalizované parametry pseudopotenciálů též pro „těžké atomy“. Statistické vyhodnocení výsledků pro 83 komplexů protein-ligand dalo pro 81% případů odhad směrodatné odchylky (RMSD) menší než 2,0 Å.

GoldMine – Program pro usnadnění zpracování výsledků získaných programem GOLD při hledání optimálního umístění skupiny ligandů v proteinu a pro vyhodnocování optimálního ligandu s volitelnými vahami jednotlivých kriterií („figures of merit“) vystupujících v účelové funkci („evaluation of ligands docking using different scoring functions“).

DASH – Software pro řešení krystalových struktur z práškových dat, tj. na základě měření práškových difraktogramů. Program využívá metodu „simulated annealing“ k hledání globálního minima účelové funkce.

O vysokém zájmu vědecké komunity o CCDC software svědčí počty pracovišť, která zakoupila licenci, a počty pracovišť spolupracujících při vývoji softwaru uvedené v následující tabulce I.

5. Příklady aplikací Cambridgeské strukturní databáze

Rozsáhlý seznam prací založených na použití Cambridgeské strukturní databáze uvedený na stránkách http://www.ccdc.cam.ac.uk/free_services/webcite/ nyní obsahuje 2304 publikací a poskytuje tedy dostatečné podklady pro získání přehledu, co vše lze s CSD zkoumat.

Pokud jde o klasické příklady využití CSD v různých oblastech technologií, chemie, fyziky a výuky, najde čtenář řadu potřebných odkazů v kapitole 7.

Abychom ukázali variabilitu možných aplikací CSD, uvedeme zde dva speciální příklady použití CSD v naší laboratoři. První příklad poukazuje na statistickou podstatu závěrů odvozovaných z CSD.

Tabulka I
Počet instalací Cambridgeské strukturní databáze, počty prodaných licencí pro specializované programy CCDC a počet pracovišť spolupracujících na vývoji systému CSD v roce 2007

Software	Akademická pracoviště	Komerční pracoviště	Spolupracující vývojová pracoviště
CSD system	1070	138	55
GOLD	282	65	13
SuperStar	48	26	1
Relibase+	25	40	10
DASH	85	14	14

Historický příklad sydnonů ukazuje, že malý počet experimentálních dat (struktur v CSD před 30 lety) dal kvalitativně správné závěry, ale pro kvantitativní hodnocení ideální korelace strukturních parametrů je vhodné analýzu opakovat vždy, pokud se podstatně zvýší počet struktur v CSD. Větší počet experimentálních dat umožňuje provést vícerozměrnou analýzu, tj. provést podrobnější analýzu vlivu dalších chemických parametrů korelovaných se zkoumaným jevem.

Druhý příklad ukazuje jiné netypické využití CSD při analýze chování polymerů v prostředí biologických makromolekul. V makromolekulární databázi PDB (cit.[5]) lze najít velký počet struktur biomakromolekul s polymery sorbovanými na povrch proteinu. Tyto interakce byly občas zpochybňovány, protože vzhledem k nízké vzájemné afinitě těchto makromolekul bývají mapy elektronové hustoty v těchto místech difusní a polymer bývá zastoupen s nižší pravděpodobností výskytu („occupancy factor“). Přestože CSD neobsahuje ani polymery ani biomakromolekuly, hraje zde CSD velmi významnou úlohu. Obsahuje totiž přesně a spolehlivě stanovené vzory (templáty), které simulují interakce mezi polymery a proteiny, a ty slouží jako bezpečná indikace pro verifikaci nebo zamítnutí možných strukturních modelů pro interakce mezi rafinovanými proteiny a polymery.

Ve třetí části jsou uvedeny odkazy na instruktáž pro využití CSD v obecné chemii, které CCDC považuje za vhodné pro výuku studentů na vyšších středních a vysokých školách.

5.1. Vlastnosti fotochromních sloučenin obsahujících sydnonový kruh – indukční a mezomerní efekt

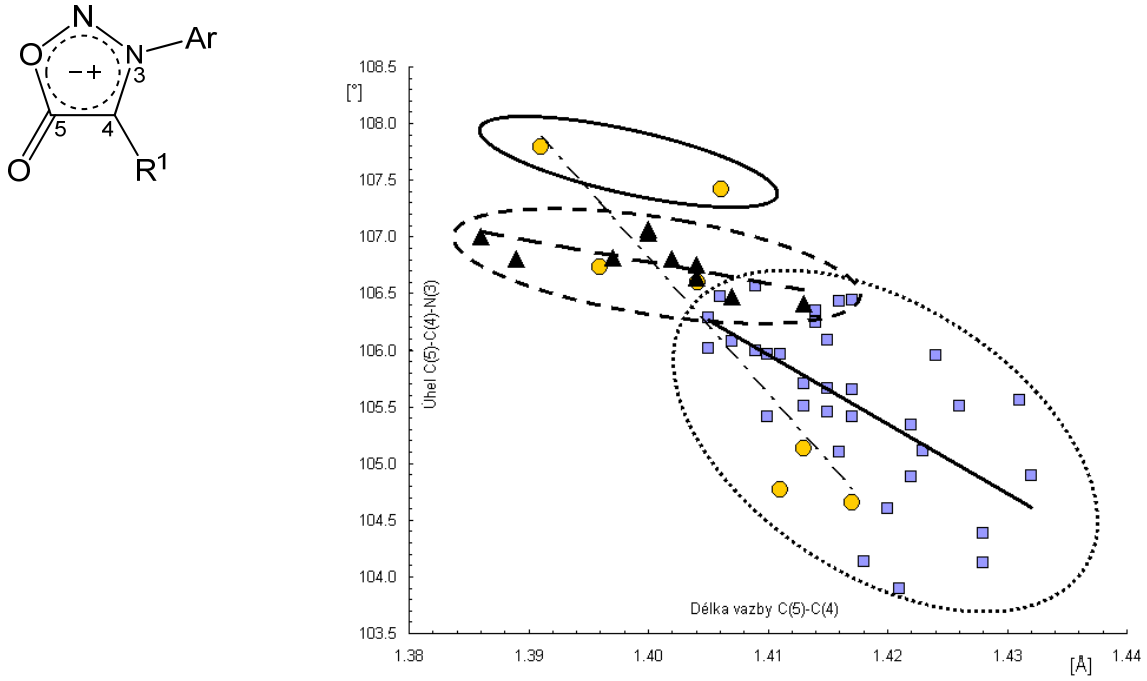
V následujícím příkladu se pokusíme ukázat, že příliš jednoduchá analýza – například pouhé porovnání geometrie dvou různě substituovaných molekul – může vést k nepřesným závěrům. Při teoretickém studiu molekulární struktury totiž často požadujeme odvození závislosti jednoho strukturního parametru na jiném zvoleném parametru. V realitě však každá změna molekuly produkuje jisté změny všech do jisté míry vzájemně korelovaných strukturních parametrů. Parametry nezahrnuté do analýzy se nám v projekci do dvojdimenzionálního grafu jeví jako poruchy, které působí, že experimentální body neleží přesně na očekávané křivce závislosti. Rozptyl parametrů, který při zpracování pozorujeme, tedy obvykle není způsoben nepřesností experimentu*, ale může vyplývat ze způsobu prováděné analýzy.

Pokud by distribuce „chyb z vynechaných parametrů“ byla symetrická, byla by proložená křivka správná. Další problém, zřejmý i v tomto příkladě, je, že výběr struktur řešených v laboratořích a deponovaných v CSD není náhodný. Některé substituenty se vyskytují velice často, jiné zcela chybí. Přesná a spolehlivá analýza tedy vyžaduje alespoň základní statistické vyhodnocení a hlubší zamyšlení. Poměrně náročná mnoharozměrná analýza je zde vždy vítaná.

Molekulární struktury s různě substituovaným sydnonovým kruhem mají velmi odlišné fotochromní vlastnosti (změny absorpčních spekter vlivem ozáření) měnící se s chemickými vlastnostmi substituentů[7–10].

Sydnony patří mezi mezoiontové sloučeniny s různým podílem rezonančních struktur, a proto je zde předpoklad, že substituenty v poloze 3 a 4 sydnonového kruhu svým σ-elektron indukčním (–I,+I) efektem nebo mezomerním efektem (–M, +M) ovlivní rozložení náboje, stupeň konjugace systému, a tedy i experimentálně pozorovanou geometrii molekuly. Na otázku, jak se mění geometrie centrálního sydnonového kruhu se záměnou substituentu v poloze 3, odpovídá graf na obr. 1, v němž jsou vyneseny pozorované hodnoty valenčního úhlu C5-C4-N3 proti délce vazby C5-C4 pro měřené struktury sydnonů s různými substituenty. Sloučeniny vybrané ke studiu mají aromatický substituent v poloze 3 a různé substituenty v poloze 4.

* Přesnost molekulární struktury získaná rtg difrakcí (v jednotkách 0,001 Å; 0,01°) je obvykle o řád větší než rozptyl bodů pozorovaných na jednoduchých dvojdimenzionálních grafech při statistické analýze CSD (v jednotkách 0,01 Å; 0,1°).



Obr. 1. **Změny geometrie sydnonového kruhu vyvolané substituentem R^1 v poloze 4.** Konkrétně jde o scatterogram valenčního úhlu C_5-C_4-N_3 vyneseného proti délce vazby C_5–C_4 sydnonového kruhu pro 50 struktur různě substituovaných sydnonů. Struktury vytváří tři zřetelně oddělené skupiny. Horní elipsa zahrnuje dvě struktury s chlorovým substituentem, R^1=Cl. Tyto struktury mají největší úhel C_5-C_4-N_3 odpovídají negativnímu elektron-extrahujícímu-efektu. Prostřední elipsa (čárkovaná čára) označuje devět struktur bez substituentu, R^1=H, označených plným trojúhelníkem. Dlouze čárkovaná čára ukazuje lineární závislost diskutovaných parametrů u devíti struktur bez substituentu v poloze 4, tedy R^1=H. Dolní elipsa (tečkovaná čára) označuje struktury, u nichž je v poloze 4 sydnonového kruhu substituent s alternující dvojnou vazbou. Plná čára ukazuje lineární závislost diskutovaných parametrů u struktur s alternující dvojnou vazbou v substituentu R^1 vykazující mezomerní resonanční efekt. Velký rozptyl bodů kolem přímky proložené metodou nejmenších čtverců svědčí o tom, že sledované strukturní parametry jsou u těchto částečně konjugovaných systémů vysoce korelované s dalšími parametry. Je to indikace pro použití vícerozměrné analýzy a vyhledání ortonormálních souřadnic popisujících lépe chování systému. Sedm bodů označených kolečky odpovídá strukturám dostupným v roce 1980. Strmá, krátce čárkovaná čára ukazuje lineární závislost mezi oběma diskutovanými parametry odvozenou z malého počtu tehdy známých dat.

V roce 1980, když jsme publikovali sérii struktur různě substituovaných sydnonů, bylo známo pouze 7 struktur splňujících požadované vlastnosti. Jim odpovídá sedm bodů na grafu označených kolečky. Graf ukazuje, že záporný náboj substituentu R^1 = Cl (dva nejvýše ležící body v grafu) vyvolává zvětšení úhlu C5-C4-N3 cca o 1° a zmenšení vazebné vzdálenosti C5-C4 cca o 0,01 Å. Alternující dvojná vazba u substituentu v poloze 4 (tři nejníže ležící kolečka v grafu na obr. 1) vede ke zmenšení úhlu C5-C4-N3 až o 3° a zvětšení vazebné vzdálenosti C5-C4 až o 0,03 Å. Graf lineární závislosti proložený metodou nejmenších čtverců (krátce čárkovaná čára) vykazuje silnou korelaci obou sledovaných parametrů.

Po roce 2000 se zájem o různě substituované sloučeniny odvozené od sydnonového kruhu výrazně oživil. V současné době je v CSD 50 struktur s různými kombinacemi substituentů v polohách 3 a 4 splňujících naše požadavky. Graf doplněný o tyto struktury ukazuje, že výše uvedené závěry odvozené z malého počtu experimentálních bodů byly v zásadě správné, ale obdobný graf, nyní s 8krát větším počtem experimentálních bodů, dává podstatně více informací (viz legenda pod obr. 1). Evidentně, větší počet experimentálních dat dává kvalitativně lepší možnosti extrahovat hledané obecné závislosti. Při analýze je proto vždy vhodné vycházet z obecněji formulovaného problému korespondujícího s velkým počtem struktur v PDB a teprve po analýze pozorovaných trendů a objasnění souvislostí problém zúžit na pouze na jednotlivý analyzovaný případ.

Velký počet struktur dostupných v současné době přesvědčivě ukazuje, že u substituentů s alternujícími dvojnými vazbami v poloze 4 často dochází k částečné konjugaci se sydnonovým kruhem. V pevné fázi se torzní úhel natočení substituentů v poloze 4 umožňujících konjugaci pohybuje 23krát v rozmezí 0–35º a 9krát v rozmezí 35–60º (medián všech dat 21º).

Naopak aromatické substituenty v poloze 3 jsou v pevné fázi vždy silně vytočeny vůči rovině sydnonového kruhu a ke konjugaci prakticky nedochází. Torzní úhel ukazující pootočení aromatického substituentu v poloze 3 se pohybuje v intervalu 35–90º (medián 63º). Torzní úhel v rozmezí 35–60º je u 13 sloučenin, a v rozmezí 60–90º u 19 sloučenin.

5.2. Interakce polyethylenoxidu s proteiny

V Proteinové strukturní databázi[5], tj. v krystalových strukturách biologických makromolekul, jsou občas pozorovány fragmenty polymerů adsorbované na některých typických místech povrchu proteinu. Přesto je sorpce polymerů na povrchu biologických makromolekul často zpochybňována pro neostrost elektronové hustoty v těchto místech, časté nižší okupanci** a také z toho důvodu, že lokalizace neznámých ligandů v makromolekulárních databázích je vždy zatížená jistou mírou nejistoty. Problematika interakce polymerů s proteiny v atomárním rozlišení nebyla dosud systematicky zkoumána. Cambridgeská strukturní databáze nám umožnila potvrdit věrohodnost většiny těchto pozorování porovnáním chování nízkomolekulárních a vysokomolekulárních ligandů. Tato studie vycházející z CSD ve svých důsledcích umožnila nový pohled na procesy řídící krystalizaci biologických makromolekul[22].

Prvním krokem bylo nalezení interakcí mezi oligomery a bioanalogickými molekulami v Cambridgeské strukturní databázi. Pokud je geometrie interakce polymeru v proteinové strukturní databázi ve shodě s konfiguracemi zjištěnými v CSD, můžeme považovat interakci protein-polymer v makromolekulární databázi PDB za oprávněnou. Takto tedy lze pomocí spolupráce mezi třemi strukturními databázemi CSD, PDB a PolyBase obejít fakt nižšího experimentálního rozlišení v případě biologických makromolekul.

Poměrně obsáhlé experimentální výsledky o pozorovaných interakcích mezi polyethylenoxidem a typickým vodorozpustným proteinem získané z CSD a PDB můžeme po značných zjednodušeních zobecnit takto:

Pokud není ve sledovaném roztoku kompetitivnější ligand, bude se polyethylenoxid snažit vytvořit následující mezimolekulární interakce:

5.2.1. Interakce polyetheru s ligandem nesoucím kladný náboj

CSD. Ve strukturní databázi organických molekul vidíme velké množství příkladů interakcí lineárních segmentů obsahujících oligoethylenoxidový fragment a nebo vidíme předformované makrocyklické polyethery tvořící 4 až 10 interakcí etherických kyslíků s touž kladně nabitou skupinou. Statistika výskytu jasně vykazuje preferenci 5–6 etherických kyslíkových atomů sbalených kolem kladného iontu ve vzdálenosti závislé na jeho iontovém poloměru (například nitrobenzo-18-crown-6.RbSCN)[11].

PDB. Obdobnou experimentálně ověřenou interakci polyetherů sorbovaných na povrchu proteinů vidíme též v proteinové strukturní databance. Nad povrchem proteinu v místě povrchově exponovaného kladného náboje (Lys, Arg, His) vytváří polyetherový řetězec kruh etherických kyslíkových atomů vzdálených 3–4 Å od atomu dusíku nesoucího v tomto případě kladný náboj na povrchu proteinu[14]. Po dobu sorpce polymeru se povrch proteinu v tomto místě virtuálně změní z hydrofilní oblasti nesoucí kladný náboj na téměř 100 Å velikou hydrofobní plochu efektivně neutralizující vliv pozitivního náboje na okolí. Tento fakt nám umožňuje efektivně ovlivňovat vlastnosti proteinů a stavíme na něm například návrh nových krystalizačních postupů nezbytných pro experimentální řešení struktury[13,15].

Povrch proteinu a lineární charakter polymerů pozorovaných v PDB brání polymeru zaujmout ideální konformaci polyetherového kruhu, kterou vidíme v CSD, a proto bývá afinita mnohonásobně nižší. V optimálním případě jsme dostávali při výpočtu energií 2–8krát nižší hodnoty stabilizační energie komplexu v případě vazby na protein. Tomu odpovídá skutečnost, že vazba polymerů na proteiny nemusí být v čase stabilní. Proto je polymer často rafinován s nižším zastoupením („occupancy factor“) a v mnoha případech jistě není sorpce polymeru na protein difrakčními prostředky vůbec identifikována (pokud pravděpodobnost jeho výskytu v daném místě, vyjádřená pomocí „occupancy factor“ klesne pod 0,2).

5.2.2. Interakce polyetheru s více ligandy nesoucími kladný náboj

CSD. U makrocyklických polyetherů s kruhem obsahujícím alespoň 24 atomů pozorujeme v CSD často koordinaci ke dvěma kationtům a nebo obklopení kationtu prohnutou elipsou etherických kyslíkových atomů. K úplnému obalení kationu dochází až od kruhů obsahujících 30 atomů (například dibenzo-30-crown-10)[12].

PDB. V PDB se zdá být obalení dvou kladně nabitých skupin umístěných blízko sebe jedním řetězcem polyetheru značně nepravděpodobné vzhledem k nutnosti přizpůsobení se tvaru povrchu proteinu a skutečně takovéto interakce nebyly v PDB nalezeny.

5.2.3. Tvorba vodíkových můstků

CSD. CSD skýtá množství informací o interakcích oligomerů s neutrálními molekulami obsahujícími donory vodíkových vazeb.

PDB. V PDB jsou takové interakce často pozorované

** V experimentu pozorujeme reálný stav molekuly za daných podmínek (teplota, tlak, stav okolních molekul, rozpouštědlo, pH, atd.). Stává se tedy, že pozorujeme smíšený stav systému (například určitá funkční skupina se v krystalu vyskytuje ve dvou různých konformacích). Pozorovaná el. hustota potom odpovídá váženému součtu elektronových hustot pro oba stavy. Zastoupení popisovaného stavu („occupancy factor“) udává podíl popisované konformace na celkovém počtu molekul.

ve strukturách proteinů v dobrém rozlišení. Nacházíme zde fragmenty polymerních řetězců (zpravidla kratší než 60 atomů) nekovalentně vázané na povrch proteinu. Polymerní řetězec prochází obvykle podél zářezů tvořených smyčkami nebo doménami proteinu bohatých na donory vodíkových vazeb. Velice flexibilní řetězec polyethylenoxidu dovede poměrně úspěšně kompenzovat entropickou i enthalpickou ztrátu energie vyvolanou doprovodným vytlačením molekul vody z tohoto zářezu. Přesto tato vazba polyethylenoxidu na protein bývá energeticky slabá. Na koncích těchto segmentů, kde již polymer nenachází vhodné rozložení donorů vodíkových vazeb, se polymerní řetězec uvolňuje od povrchu proteinu a zůstává rozpuštěn v blízkosti povrchu. Je pravděpodobné, ovšem ne zaručené, že další segmenty polymerů pozorované na povrchu molekuly proteinu patří k téže molekule polymeru.

5.2.4. Hydrofobní kontakty

CSD. V CSD je možné nalézt mnoho příkladů úplného obalení kationu deseti etherickými kyslíkovými atomy izolujícími dokonale kation od okolí [12].

PDB. Také v PDB existují konfigurace, kde solvatovaný kladný iont je v roztoku obalen molekulou polyetylenoxidu a celý tento komplex dosedá svým hydrofobním obalem na hydrofobní segment povrchu proteinu. Přestože z energetického hlediska tato konfigurace může být stabilní, přesná pozice polymerního řetězce zde není energeticky stabilizována a proto má rtg strukturní analýza určující časově a prostorově zprůměrovanou strukturu problémy tuto konfiguraci zobrazit.

6. Dostupnost CSD v ČR

Cambridgeská strukturní databáze je provozována obvykle jako lokální instalace pod systémem MS Windows licencovaná pro jednu IP adresu (objednávky na adrese: hasek@imc.cas.cz). Internetový přístup pro české uživatele je možný po registraci ve Fyzikálním ústavu AV ČR (kontakt: dusek@fzu.cz).

Nadstavbové programy GOLD, DASH, RELIBASE+ a SUPERSTAR jsou distribuovány přímo administrativním centrem v Cambridge pouze jako lokální licence (http://www.ccdc.cam.ac.uk/contact/obtaining_products/). Od všech produktů je možné vyžádat si časově omezenou zkušební verzi „free evaluation copy“.

Licence na používání plné verze Cambridgeské strukturní databáze je placena jednou ročně. Nový software a data jsou dodávány v jarních měsících na DVD a nově přibývající data je možné doplňovat čtvrtletním stahováním doplňků ze serveru http://www.ccdc.cam.ac.uk. Plnou verzi CSD v současné době provozují FZU AV ČR, VŠCHT, ÚMCH AV ČR, ÚOCHB AV ČR, PřF UK, MU Brno, VFU Brno, Univerzita Pardubice. Licence pro české uživatele se vztahují pouze na nekomerční využití a nekomerční uživatele.

7. Školní verze Cambridgeské strukturní databáze

Čtenářům, kteří se chtějí seznámit s jednoduchými možnostmi nabízenými CCDC, doporučuji návštěvu stránek http://webcsd.ccdc.cam.ac.uk/teaching_database_demo.php. Bez nutnosti jakékoliv instalace a zcela bezplatně si zde každý může velmi snadno vyzkoušet základní manipulace s molekulární strukturou. Návštěva těchto stránek nevyžaduje žádné speciální znalosti a může být součástí výuky chemie na všech základních a středních školách.

Při prohlížení stránek, které se automaticky otevřou, si vyzkoušejte rotace a změny měřítka zobrazené struktury pomocí myši, různé způsoby zobrazení atomů i molekulárních povrchů, projděte několik struktur zatrháváním referenčních kódů na liště na levém okraji. Nezapomeňte kliknutím na „Launch External Viewer“ zobrazit strukturu v lepším prohlížeči MERCURY (možno zdarma stáhnout a nainstalovat). Na horní liště zvolte v menu „Databases“ položku „Open the Teaching Subset“. Tím získáte možnost pracovat s pěti sty vybranými molekulárními strukturami. Dole na liště označte „Short Contact“. Postupným klikáním na červené kuličky na konci naznačených vazeb k sousedním molekulám a pootáčením obrazu myší získáte přehled o tom, jak molekuly vzájemně interagují a vytvářejí pevnou fázi.

Připravené příklady využití CSD pro prezentace studentům chemie jsou dostupné na adrese http://www.ccdc.cam.ac.uk/free_services/teaching/ a zahrnují:

- sledování změn geometrie aromatických systémů vlivem různých substituentů,
- příklady a vysvětlení „valence-shell-electron-pair-repulsion (VSEPR)“,
- příklady a vysvětlení pojmu hapticity u koordinačních sloučenin,
- příklady a vysvětlení konformačního chování cyklických molekul,
- příklady a vysvětlení izomerie, molekulární a krystalové chirality,
- příklad analýzy reakčních intermediátů.

Mnoho dalších příkladů aplikací Cambridgeské strukturní databáze v různých vědních oborech lze nalézt seznamu literatury na adrese http://www.ccdc.cam.ac.uk/free_services/webcite/ a v níže uvedené literatuře[16–21].

Poněkud náročnější je bezplatná „školní verze CSD“, která po instalaci umožňuje demonstrovat základní rysy vyhledávacího systému **ConQuest** na libovolném počtu počítačů v dané instituci. Omezený vzorek 500 struktur je vybrán z CSD tak, aby zahrnul nejčastěji probírané sloučeniny ve školních učebnicích. Popisovanou školní verzi CSD lze stáhnout zdarma z adresy http://www.ccdc.cam.ac.uk/free_services/teaching/. K samostatnému prohlížení molekulárních a krystalových struktur je ještě třeba stáhnout školní verzi programu MERCURY z adresy http://www.ccdc.cam.ac.uk/free_services/mercury/.

Pokud má daná instituce placenou verzi CSD alespoň pro jednu IP adresu, může si vytvořit svoji vlastní databázi 500 struktur vybraných podle potřeb a zaměření pracoviště a používat tuto svoji malou databázi pro výuku na libovolném počtu počítačů. CSD manuály[16–21] je možné zdarma stáhnout nebo v tištěné formě zakoupit prostřednictvím www CCDC.

Tento projekt je podporován v rámci grantů IAA500500701 a 305/07/1073.

LITERATURA

1. Allen F. H., Motherwell W. D. S.: Acta Crystallogr. *B 58*, 407 (2002).
2. van de Streek J.: Acta Crystallogr. *B62*, 567 (2006).
3. Orpen A. G.: Acta Crystallogr. *B58*, 398 (2002).
4. Taylor R.: Acta Crystallogr. *D58*, 879 (2002).
5. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E.: Nucleic Acids Res. *28,* 235 (2000).
6. Hašek J., Dohnálek J., Skálová T., Dušková J., Štěpánková A., Kolenko P.: Polymer Structure Database. In: *Proceedings of the EPDIC 2010, Darmstadt 2010*.
7. Hašek J., Obrda J., Huml K., Nešpůrek S., Šorm M.: Acta Crystallogr. *B35*, 437 (1979).
8. Hašek J., Obrda J., Huml K., Nešpůrek S., Šorm M.: Acta Crystallogr. *B35*, 2449 (1979).
9. Hašek J., Obrda J., Huml K., Nešpůrek S., Chojnacki H., Šorm M.: Acta Crystallogr. *B34*, 2756 (1978).
10. Hašek J., Beurskens P. T., Obrda J., Nešpůrek S., Schenk H., Goubitz K., Huml K., Schagen J. D.: Collect. Czech. Chem. Commun. *50*, 1764 (1985).
11. Hlavatá D., Hašek J., Huml K.: Acta Crystallogr. *B 34*, 416 (1978).
12. Hašek J., Hlavatá D., Huml K.: Acta Crystallogr. *B 36*, 1782 (1980).
13. Hašek J.: Z. Kristallogr. Suppl. *23*, 613 (2006).
14. Skálová T., Dohnálek J., Ostergaard L. H., Ostergaard P. R., Kolenko P., Dušková J., Štěpánková A., Hašek J.: J. Mol. Biol. *385*, 1165 (2009).
15. Skálová T., Dušková J., Hašek J., Kolenko P., Štěpánková A., Dohnálek J.: J. Appl. Crystallogr. *43*, 737 (2010).
16. CCSD manual "Using the Cambridge Structural Database for Teaching", CCDC Cambridge 2008.
17. CCSD manual "CSD Workshop", 76 pp., CCDC Cambridge 2001.
18. CCSD manual "Cambridge Structural Database System. Worked Examples", CCDC Cambridge 2001.
19. CCSD manual "Mogul 1.0 User Guide", CCDC Cambridge 2004.
20. CCSD manual "Conquest 1.7 User Guide", CCDC Cambridge 2004.
21. CCSD manual "Conquest 1.7 Tutorials", CCDC Cambridge 2004.
22. Hašek J.: J. Synchrotron Radiation *18*, 50 (2011).

J. Hašek *(Institute of Macromolecular Chemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic):* **Cambridge Structure Database of Organic and Organometallic Compounds**

The paper describes history and present state of the Cambridge Structure Database of Organic and Organometallic Structures (CSD) and the related software. The reader finds references to the CSD applications in chemistry and biochemistry, and to the user manuals. Two special examples from the author's laboratory give a glimpse at possibilities and the expected reliability when forecasting general trends from experimentally determined structures. The first example shows how the information content increases with growing number of experimentally determined structures. The analysis made in 1980 on seven differently substituted sydnone rings showed only significant changes of geometry as an influence of substituent charge and/or conjugation between the sydnone ring and its aromatic substituents. Nowadays, the same analysis made after 30 years on 50 differently substituted sydnone rings available in the CSD shows a necessity of clustering, gives more details and also more exact correlations between chemical composition and molecular geometry. The second example shows an extrapolation of experimental information on organic compounds extracted from the CSD to a different field of science – the analysis of protein -polymer interactions. Information on accessibility and usage of the CSD in the Czech Republic closes the paper.