# Using the Cambridge Structure Database of Organic and Organometalic Compounds in Structure Biology

**Jindřich Hašek**

*Institute of Macromolecular Chemistry, Academy of Sciences of the Czech Republic, Heyrovského nám. 2, Praha 6, Czech Republic*
*hasek@imc.cas.cz*

## Introduction

Experimentally determined structures are collectred in several separate databases. Macromolecular structures of biological interest are collected in the Protein Data Bank (PDB) /5/, polymer structures in PolyBase /6/, inorganic structures in Database of Inorganic Crystal Structures (http://www.fiz-karlsruhe.de/icsd_content.html), structures of metals and alloys in CRYSMET@ database (http://www.tothcanada.com/databases.htm). Organic structures of biological interest discussed in this paper are collected in the Cambridge Structure Database (CSD).

## Cambridge Structure Database

The CSD maintained by the Cambridge Crystallographic Data Center (CCDC) contains more than half milion of organic and organometalic compounds [1,2]. Most structures deposited in the database are determined experimentally by diffraction methods with high accuracy. Coordinates of atoms have standard deviations usually in the range 0.001 – 0.01 Å (i.e. better than 1 picometer). With this accuracy, the CSD is useful source of information on bond strengths, ionization states, etc. It provides also a reliable information on electron density of hydrogens that is usually not available in the PDB (Fig.1).

About hundred thousand compounds deposited in the CSD are of biological relevance. One can find here many drug leads, peptides, biologically active compounds, etc. About six thousands compounds of biological origin are labeled as natural products.

The CSD is used at more than a thousand universities from 69 countries and at about 200 commercial organizations from 15 countries. Most installations are in the USA (243), Japan (105), France (99), Germany (69), Russia (59), Poland (55), Spain (51), Italy (40), GB (37), Canada (35). The Czech Republic belongs to the first five states with permanent subscription to the CSD since 1972. Non-commercial users from the Czech Republic can use the CSD via registration at the Institute of Physics of the Academy of Sciences in Praha (contact: dusek@fzu.cz) or can install their local versions at their own personal computers (contact: hasek@imc.cas.cz). A basic information on the database [3,4], on the related software, and on free services is available at the address http://www.ccdc.cam.ac.uk/products/.

## Applications

Searching for structural properties of groups of the compounds extracted from CSD and using statistical tools one should remember that the contents of database does not cover the space of compounds uniformly, i.e. the extracted set of structures is not a statistically random sample. For example in peptide analysis, the conformationally interesting aminoacid residues (AA) are represented much more frequently in CSD than other residues. There are several hundreds of peptides with prolin or glycin in CSD because an interest of people has been concentrated especially to the structural effects of different sequences containing these two aminoacids. Other AAs are found much less frequently.

Another example is a limited crystallizability of long peptides. Short peptides have usually well defined conformations and can be crystallized easily. Longer peptides with 10–25 AA are already fully hydrated (Fig.2), but still they are not long enough to form a stable 3D structure in solution. As a result, they are conformationally unstable, their crystallization is very difficult and therefore one can find only a few structures in the CSD.

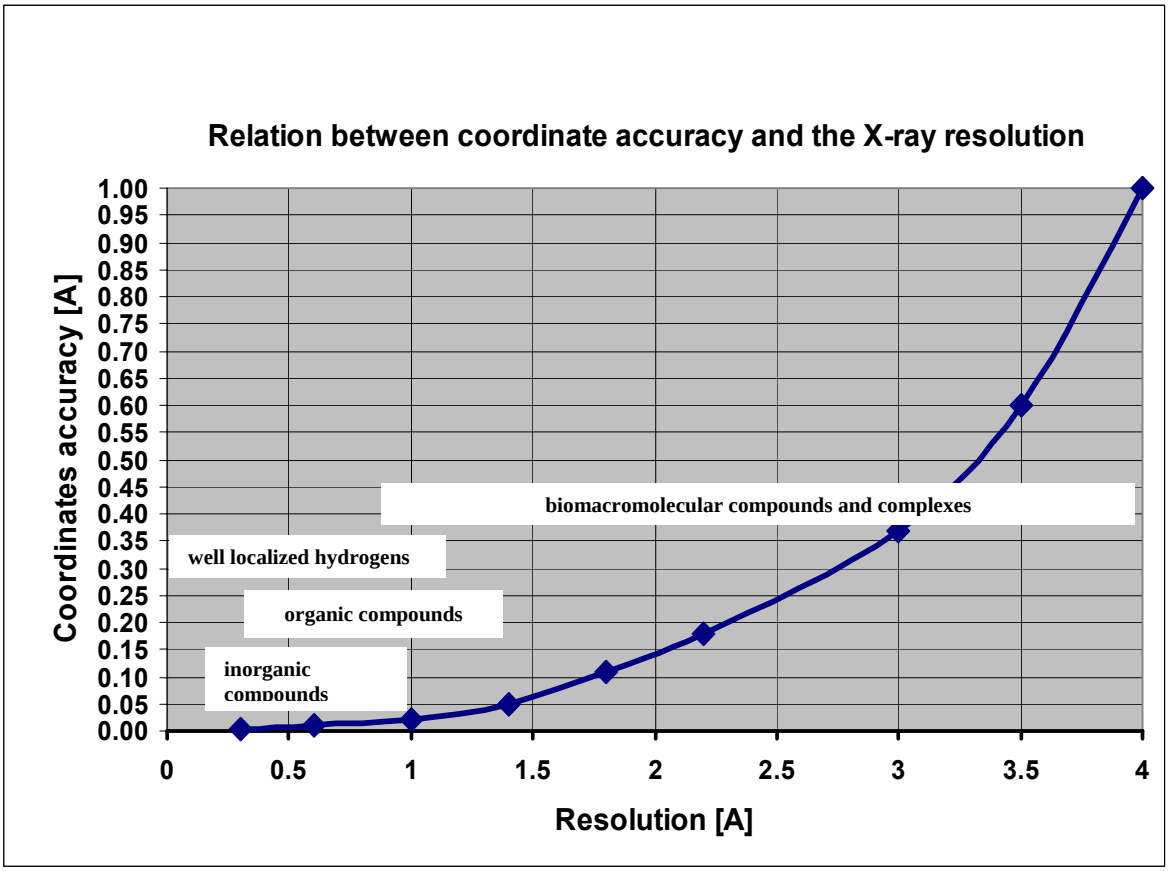


Fig.1. Relation between the diffraction limit (resolution) and the estimated standard deviation of atomic positions for structures determined by diffraction methods ( 1 Å = 100 pm ).

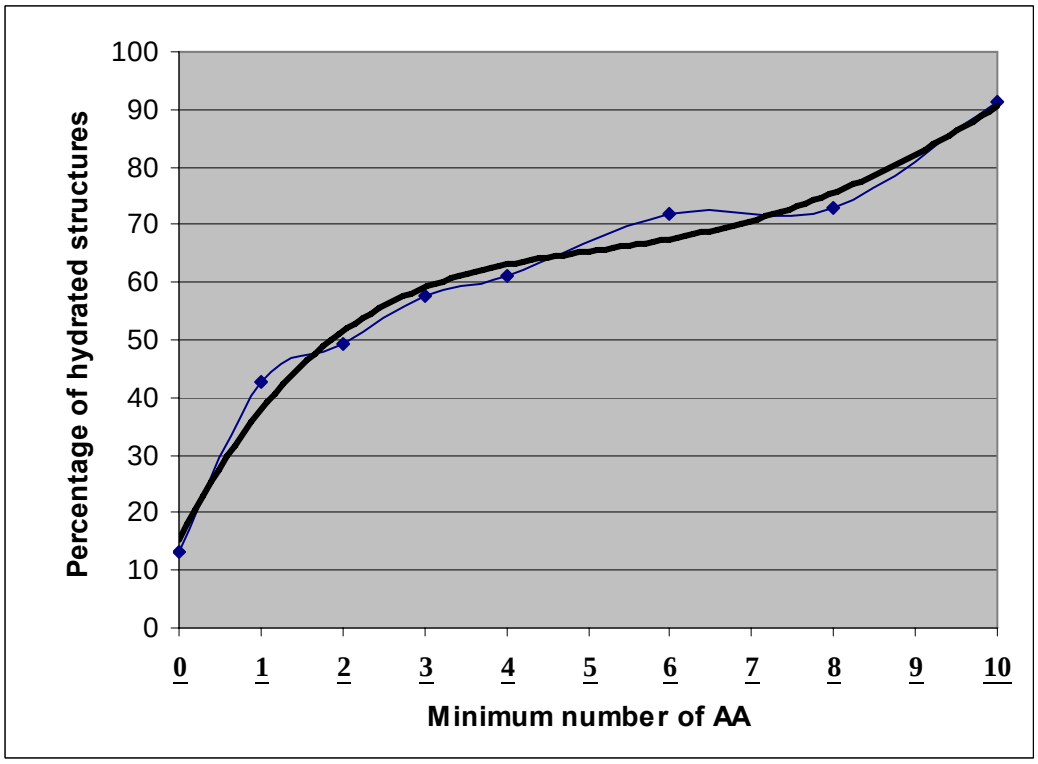


Fig.2. Probability of hydrated crystals increases with the number of aminoacid residues.

CCDC provides also software for analysis of the Protein Data Bank (PDB) of experimentally determined bio-macromolecular structures /5/. The efficiency and accuracy of structure determination by X-ray crystallography depends mainly on the diffraction quality of crystals (Fig.1) and in principle has no limitation as for the complexity of the macromolecular complexes. Thus, a search for 3D fragments and analysis of these large macromolecular complexes with milions of atoms is demanding. Software “Relibase” provided by CCDC is very useful for 3D scanning of the PDB contents and the program “GOLD” is designed for empirically based 3D ligand docking interesting for special interest groups of structure biologists and drug designers. The synthetic polymers collected in the Polymer Structure Database (PolyBase) [6] can be analyzed by CCDC program MERCURY.

## Conclusion

The CSD is available for non-commecial users in the Czech Republic for a small license fee payed each year. The related software for extension of structure analysis based on the CSD and PDB data can be purchased separately. A free evaluation copy can be requested for any CCDC product at the address http://www.ccdc.cam.ac.uk/support/product_references/. For teaching purposes, one should register via internet for a free version of CSD contaning 500 structures http://www.ccdc.cam.ac.uk/free_services/teaching/.

The activities reported in this paper are supported partly from the GA AS project IAA500500701 and the GA CR project 305/07/1073.

1. Allen F.H., Motherwell W.D.S. Acta Crystallogr. B 58, 407-422, (2002).
2. Taylor R., Acta Crystallogr. D58, 879-888, (2002).
3. CCSD manual "Using the Cambridge Structural Database for Teaching", 103 pp., CCDC Cambridge (2008).
4. CCSD manual "Conquest 1.7 User Guide", 226 pp., CCDC Cambridge (2004).
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, 235-242, (2000).
6. Hašek J., Labský J.: Database of polymer structures - PolyBase, CSCA Praha (1995).