



Main Lectures - Session V, Wednesday, June 13

L9

SYSTEMATIC ANALYSIS OF CRYSTAL AND MOLECULAR STRUCTURES

J. Hašek, J. Dohnálek

*Institute of Macromolecular Chemistry AV ČR, Heyrovského nám.2, 16206 Praha 6,
hasek@imc.cas.cz*

Currently available sources of experimentally determined molecular structures for analysis of relations between structure and function are:

Cambridge Structure Database of Organic and Organometallic Compounds (CSD) [1] is distributed in our country by the Crystallographic Association. It contains presently 703 297 organic structures, 339 297 containing at least one metal atom and 221 037 containing at least one halogen atom.

The database is searched by self-explanatory program ConQuest. The best tool for visual inspection of structures in crystalline phase is MERCURY 3.0 [2]. Some other aspects of using the CSD are reviewed in [3]. Teaching modules are at http://www.ccdc.cam.ac.uk/support/documentation/csd/teaching_egs/toc.html. Advanced case studies are at http://www.ccdc.cam.ac.uk/case_studies/small_mol/ for inorganic structures and at http://www.ccdc.cam.ac.uk/case_studies/life_science/ for life sciences. The teaching version is free of charge. The full version is available for a reasonable fee paid yearly.

Protein Databank (PDB) [4] contains 81 553 macromolecular structures mostly of biological origin. 71 508 of them were determined by X-ray diffraction, 9 354 by solution NMR, 51 by solid state NMR, 38 by neutron diffraction, 32 by electron diffraction, and 428 by electron microscopy.

75 525 of these structures are proteins, 1 340 are DNA fragments, and 907 RNA. 41 052 structures originate from eukaryota, 29 945 from bacteria, 4 826 from viruses, 3 215 from archaea, and 3 743 from other sources. As far as eukaryota, 19 952 structures of biological macromolecules originate from *Homo Sapiens*, 5 854 from *E-coli*, 3 513 from *Mus Musculus*, 2 148 from *Saccharomyces Cerevisiae*, 2 047 from *Bos Taurus*, etc. The most frequently present enzyme types are hydrolases 16 952, transferases 12 493, oxydoreductases 7 480, lyases 3 340, isomerases 1 900 and ligases 1 696.

The PDB site refers to many useful programs for structure analysis, e.g. for inspection of adsorption of ligands on protein surface is the Protein Explorer [5]. The use of database is free.

The database of human disease-related VIS (<http://www.scbt.org/dbmi/drvis>) collects and maintains human disease-related VIS data, including characteristics of the malignant diseases, chromosome region, genomic position and viral–host junction sequence. The current database covers about 600 natural VIS of 5 oncogenic viruses representing 11 diseases. Among them, about 200 VIS have viral–host junction sequence.

Polymer Structure Database (PolyBase) [6] contains structures of synthetic polymers determined experimentally in solid state and several thousands of “snapshots” of the PEG-like polymers built in the environment of biological macromolecules. Non-commercial users in the Czech and Slovak Republics can get access to these services.

The Nucleic Acids Data Bank (<http://ndbserver.rutgers.edu/>) contains 5 897 structures of oligonucleotides. The link <http://ndbserver.rutgers.edu/education/index.html> refers to several useful links to teaching and educational resources. All events related to the Nucleic Acids Data Bank are archived in the NDB Newsletter http://ndbserver.rutgers.edu/NDB_news/index.html. The use of the database is free.

The Inorganic Crystal Structure Database (http://www.fiz-karlsruhe.de/icsd_content.html) currently contains 150 042 inorganic structures. 1 616 crystal structures of elements, 28 354 structures of binary compounds, 55 436 ternary compounds and 54 144 quaternary and quinary compounds. The database is updated twice a year, adding each time about 3 500 new records. The database should be purchased directly from <http://www.fiz-karlsruhe.de>.

The database CRYSTMET® - Metals and Alloys (<http://www.tothcanada.com/>) contains 139 058 entries with critically evaluated crystallographic data, atomic coordinates and calculated powder diffraction patterns for metals, alloys, intermetallics and minerals. The database is closely related to a software package for materials modelling “Materials Toolkit 2.7.1” published in [7].

The Crystallography Open Database – COD (<http://www.crystallography.net>) collects all known „small molecule” and “small-medium sized unit cell” crystal structures and makes them available freely on the internet. It contains more than 150 000 structures [8].

Simple use of databases makes usually no problem. An advanced usage may be more sophisticated and requires some experience. In case of interest in the advanced courses and discussion meetings organized in the next three years, please fill the attached tentative form. It ensures that you will receive the final information about the program and exact date of the individual courses.

The research is supported by GA ČR 310/09/1407.

1. F. H. Allen, *Acta Cryst.*, **B58**, (2002), 380-388. (www.ccdc.cam.ac.uk).
2. C. F. Macrae et al, *J. Appl. Cryst.*, **41**, (2008), 466-470.
3. J. Hašek, *Chem. Listy* **105**, (2011), 467-475.

4. H.M. Berman et al, *Nucleic Acids Res.*, **28**, (2000), 235-242. (http://ftp.wwpdb.org/pub/pdb/doc/newsletters/rcsb_pdb)
5. http://www.pdb.org/pdb/staticHelp.do?p=help/viewers/ligandExplorer_viewer.html
6. J. Hašek et al, *Zeitschrift fur Kristallogr.* **28**, (2011), 475-480.
7. Y.L. Page, J.R. Rodgers, *J. Appl. Crystallogr.* **38**, (2005), 697-705.
8. S. Grazulis et al, *Nucleic Acids Research*, **40**, (2012), 420-427.

L10

MULTIVARIATE METHODS FOR MACROMOLECULAR STRUCTURE SOLUTION FROM EXPERIMENTAL PHASES

P. Skubák, N. S. Pannu

Department of Biostructural Chemistry, Leiden University, Leiden (the Netherlands)
p.skubak@chem.leidenuniv.nl

Traditionally, the process of macromolecular structure solution consists of separate experimental phasing, density modification and model building with refinement (Fig. 1a). First, experimental phasing provides initial estimates of the phases using a probability distribution P_{ph} of several structure factors with a known relation. The initial map constructed using the initial phases is then repeatedly modified in order to introduce expected protein electron map features into the it. Since this process also introduces errors into the map, the resulting map is usually recombined with the initial map using a reciprocal space probability distribution P_{DM} , followed by another density modification iteration of the recombined map. After several iterations, the recombined map is used for iterative tracing of the model with its reciprocal space refinement using probability distribution P_{ref} .

Typically, only the experimental phasing target P_{ph} is regarded a function of several related structure factors which inherit the information about phases. The probability distributions P_{DM} and P_{ref} are considered univariate functions of a single structure factor, ignoring the experi-

mental phase information. While this approach is often sufficient, the use of the extra information may be essential at lower resolutions as it increases the low observations-to-parameters ratio or in the case of weak initial phases as it enables to retain and further improve the quality of the phase estimates.

The phase information can be indirectly incorporated into the univariate targets P_{DM} and P_{ref} , using the phase distribution obtained from phasing [1]. However, this method suffers by several theoretical and practical shortcomings. Indeed, we have shown that direct incorporation of the experimental phase information by a multivariate function for model refinement, modeling the related structure factors and the correlations between them, removes unjustified assumptions and provides higher quality models than the refinement function with indirect phase information incorporation [2]. Similarly, we have shown that a multivariate function for combination of the density modified maps and experimental maps can lead to significant improvements in quality of the electron density maps [3].

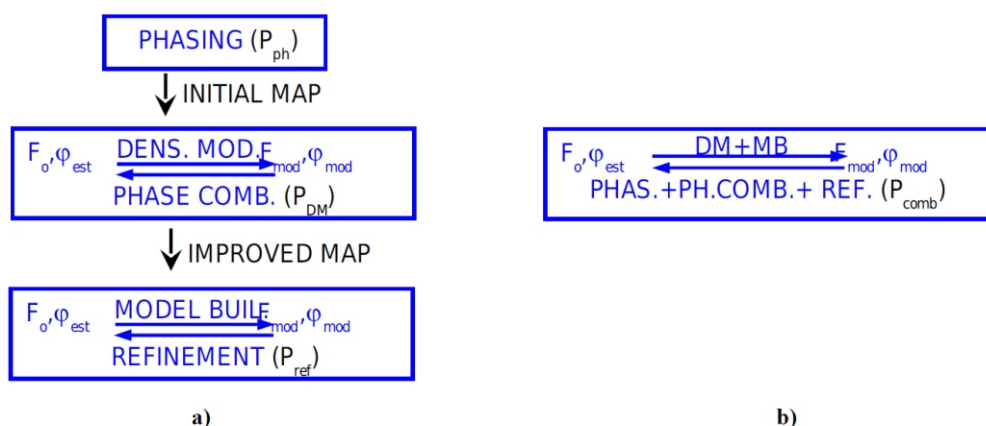


Figure 1. Process of structure solution from experimental phases.

a) The traditional process of structure solution. The steps are separated and a different function (P) is used in phasing, density modification phase combination and model refinement.

b) The new structure solution in which a single multivariate function incorporates phase information and density modification information in the current model refinement.



Recently, we have derived a novel multivariate likelihood target function for refinement of a (partial) model which incorporates both the phase information and the density modification information directly in the refinement function. The function enables to consider the traditionally separate steps of macromolecular X-ray structure determination from experimental phases as one unified process (Fig. 1b) that can lead to combined and significant improvements.

The new function has been implemented in the refinement program Refmac5 [4] and combined with density modification and model building in the CRANK suite [5]. Tests performed show the new approach allows for many structures to be built automatically that eluded current methods, especially for data at lower resolutions or with weak experimental phase information.

L11

SEMI-AUTOMATICKÉ HLEDÁNÍ LIGANDŮ V PROTEINOVÝCH STRUKTURÁCH

F. Pavelčík¹, J. Václavík²

¹Ústav chemických léčiv, ²Ústav přírodních léčiv, Farmaceutická fakulta, Veterinární a farmaceutická univerzita v Brně, Palackého 1-3, 61242 Brno
pavelcikf@vfu.cz

Úvod

Automatizované hledání ligandů v protein-ligandových komplexech je žádané v některých moderních metodách designu léčiv jako je např. FBDD (fragment-based drug design). FBDD je relativně nová metoda [1] objevování léčiv. Místo hledání celého léčiva se hledají nejprve jednotlivé malé fragmenty, které mají doplňkový povrch a vysokou interakční energii k aktivnímu místu cílového proteinu. Výběr cílových proteinů byl umožněn současným dramatickým rozvojem metod molekulární biologie, genetiky a příbuzných disciplin. Principiální průlom v designu léčiv je využití fyzikálních metod screeningu jako je NMR nebo rtg. krystalografie místo farmakologického screeningu (HTS). Princip FBDD metody je velmi jednoduchý a je zobrazen na obr. 1. Malé fragmenty se následně spojí chemickými metodami do větší molekuly, tak aby prostorové rozložení funkčních skupin bylo stejné jako ve struktuře protein-ligandových komplexů. Klasické farmakologické nebo biochemické metody obvykle vůbec nejsou schopné detekovat požadovanou biologickou aktivitu těchto malých fragmentů.

V rtg. metodách se používá koktejl 4-10 malých molekul z knihovny obsahující řádově 100 sloučenin. Pokud by byla cílová molekula složena ze tří fragmentů je taková knihovna ekvivalentní knihovně jednoho milionu testovaných sloučenin. Uvedený koktejl se aplikuje na

1. Pannu, N.S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* **D54**, (1998), 1285–1294.
2. P., Murshudov, G. N. & Pannu, N. S. *Acta Cryst.* **D60**, (2004), 2196–2201.
3. Skubak, P., Waterreus W.-J., Pannu, N.S. *Acta Cryst.*, **D66**, (2010), 783–788.
4. Murshudov, G.N., Skubak P., Lebedev A.A., Pannu N.S., Steiner R.A., Nicholls R.A., Winn M.D., Long F. and Vagin A.A. *Acta Cryst.* **D67**, (2011), 355–367.
5. Pannu, N.S., Waterreus W.J., Skubak P., Sikharulidze I., Abrahams J.P. and de Graaff RAG. *Acta Cryst.* **D67**, (2011), 331–337.

nativní proteinový krystal, který si sám vybere nejvhodnější sloučeninu. Tento proces se opakuje s jinými koktejly. Následný difrakční experiment je plně automatizovaný a celý krystalografický screening trvá pouze desítky hodin. Úlohou proteinové krystalografie je určit typ navázaného ligandu a jeho pozici v proteinové struktuře.

Metoda

Naše metoda [2] na lokalizaci malých molekul v protein-ligand komplexech, fázovaná rotační-konformační-translační funkce (PRCTF) je určitou analogií metody molekulového nahrazování (MR). Všeobecně, PRCTF je principiální nástroj na lokalizaci molekulových fragmentů v elektronové hustotě krystalu a může být formulována jako

$$F(t, \theta, \phi) = \int_V \rho_c(t) \rho_f(t, \theta, \phi) dV \quad (1)$$

kde ρ_c je elektronová hustota krystalu, definovaná pro kouli o objemu V (s poloměrem R_{ex}) v okolí mřížkového bodu, který přísluší translačnímu vektoru t . ρ_f je elektronová hustota hledaného fragmentu. Elektronová hustota fragmentu je funkcí konformace (vektor torzních úhlů, θ) a orientace (tři Eulerovy úhly, vektor ϕ). Geometrický střed fragmentu je umístěn do počátku souřadného systému.



Obrázek 1. Princip fragmentové metody designu léčiv.

Elektronovou hustotu v kouli je možné popsat pomocí sférických harmonických (Y_{lm}) a sférických Besselových [$g_l(k_n r)$] funkcí (určitá analogie atomových orbitalů, s využitím hlavního, orbitalového a magnetického kvantového čísla) a rozvojové koeficienty je možné vypočítat ze strukturních faktorů:

$$(r, \theta, \phi) = \sum_{l, m} a_{nlm} g_l(k_n r) Y_{lm}(\theta, \phi) \quad (2)$$

$$a_{nlm} = \int_0^a \int_0^\pi \int_0^{2\pi} \rho(r, \theta, \phi) Y_{lm}^*(\theta, \phi) R_{nl}(h, a) r^2 \sin \theta dr d\theta d\phi \quad (3)$$

Analogicky je možné určit rozvojové koeficienty elektronové hustoty fragmentu:

$$b_{nlm} = \int_0^a \int_0^\pi \int_0^{2\pi} \rho_f(r, \theta, \phi) S_{nlm} r^2 \sin \theta dr d\theta d\phi \quad (4)$$

Orientaci fragmentu je potom možné počítat pomocí FFT. Koeficienty C_{lmn} se vypočítají z koeficientů a_{nlm} a b_{nlm} . D_{mm} jsou Wignerovy matice pro moment hybnosti používané v kvantové mechanice.

$$R(\theta, \phi) = \sum_{l, m} c_{lm'm'} d_{m'm}^l(\theta) \exp[i(m - m')\phi] \quad (5)$$

V PRCTF se na lokalizaci fragmentů v elektronové hustotě, v skutečnosti elektronová hustota vůbec nepoužívá. Metoda je implementována v programu **NUT** [2].

Pro velmi malé fragmenty, nízké rozlišení nebo nepřesné fáze strukturních faktorů nemá překryvový integrál (1) někdy dostatečnou schopnost odlišit správná a falešná řešení. Za tímto účelem byl napsán malý program **DOK**, který počítá interakční energii mezi ligandem a proteinem v příslušné proteinové dutině (cavity). Kombinací dvou fyzikálně odlišných přístupů je možné eliminovat prakticky všechna nesprávná řešení.

Odkazy

1. W. Jahnke, D.A. Erlanson, *Fragment-based Approaches in drug Discovery*, WILEY-VCH, Weinheim, Germany, 2006.
2. F. Pavelcik, *J. Appl. Cryst.* **39**, (2006), 483-486.

Main Lectures - Session VI, Thursday, June 14

L13

POPIS DIFRAKČNÍCH PROFILŮ POMOCÍ POJMŮ MATEMATICKÉ STATISTIKY

M. Čerňanský

Fyzikální ústav AV ČR, v. v. i., Na Slovance 2, 182 21 Praha 8, Česká republika
cernan@fzu.cz

Podobnost tvaru difrakčních profilů s grafickým znázorněním funkcí hustoty pravděpodobnosti v matematické statistice vedla k formálnímu převzetí některých pojmů matematického aparátu teorie pravděpodobnosti a k jejich využití pro popis difrakčních profilů. Zejména se jednalo o momenty funkce hustoty pravděpodobnosti, které stručně, numericky popisují polohu, rozptyl, asymetrii a plochost rozložení náhodné proměnné.

Podobně tomu, se poloha difrakčního profilu, kromě polohy jeho maxima, může popisovat polohou jeho těžiště – což je první moment kolem počátku (souřadnic). Kromě šířky v poloviční výšce profilu, nebo kromě integrální šířky lze šířku difrakčního profilu popsat variancí – druhým centrálním momentem. Asymetrii, resp. plochost difrakčního profilu lze kvantitativně popsat pomocí parametrů souvisejících se třetím, resp. čtvrtým momentem.

Jednou z výhod popisu difrakčních profilů pomocí momentů je to, že pro funkce svázané relací konvoluce platí určité vztahy také pro jejich momenty. Zejména pro těžiště a variance jsou tyto vztahy velmi jednoduché.

Přitom je relace konvoluce u difrakčních profilů velmi důležitá. Vyjadřuje souvislost mezi profilem měřeným, instrumentálním a fyzikálním. Podobně tak mezi profilem fyzikálním, profilem od velikosti krystalických částic a profilem od mikrodeformací. Např. těžiště fyzikálního profilu lze určit přímo jen z těžiště profilu měřeného a z těžiště profilu instrumentálního.

Další výhodou je to, že druhý centrální moment – variance fyzikálního profilu má přímou fyzikální interpretaci ve smyslu velikosti krystalických částic a mikrodeformací. Čtvrté momenty byly použity k určení velikosti částic a mikrodeformací v případě, že je k dispozici jenom jedna difrakční linie. V tomto případě – u vyšších momentů – jsou však vztahy mezi momentami funkcí svázanými konvolucí složitější, a proto bylo navrženo použít z matematické statistiky, resp. teorie pravděpodobnosti další pojem – kumulanty (semiinvarianty). To je výhodné zvláště namísto vyšších momentů a v případě více než dvou funkcí svázaných konvolucí.

Tato práce vznikla v rámci realizace projektu Akademie věd České republiky – KAN 300100801.