Thank G-D

Intermediate report – CDB, a Database for Protein heteromeric complexes

Malki Aker and Shirly Ohanona, Dept. of Bioinformatics, the Jerusalem College of Technology, Lev

Academic Center.

Supervisions: Dr Zohar Barnett Itzhaki and Dr Moshe Amitay.

Abstract

Crystallographic structures of protein complexes are crucial to the understanding of molecular and cellular processes and may help in evaluation of protein-protein docking methods. Direct access to crystallographic structures of protein complexes can aid the development of new types of protein-protein docking benchmarks. CDB(Complex database) is a web database for heteromeric protein crystallographic complexes along with the crystallographic structures of the individual unbound proteins. This database can be accessed at http://www.jct-bioinfo.com/. In addition to the comprehensive queries available in the web database, we also provide an open access code, in order to facilitate a more flexible access to complexes and individual unbound proteins. The code for retrieving these structures from the PDB is given via GitHub at https://github.com/MosheAmitay/CDB. Both the web database and the available code can serve as a starting point for the construction of new experimental complexes sets of any type, and may help in conducting structural studies regarding protein complexes.

Introduction

Proteins perform a crucial role in living system, they function as catalysts, they transport and store other molecules such as oxygen, they provide mechanical support and immune protection, they generate movement, they transmit nerve impulses, and they control growth and differentiation[1].

Proteins can also interact with multiple proteins to form complex assemblies. The proteins within these assemblies can produce capabilities not afforded by the individual component proteins. Complex creation occurs at essential processes in the cell *i.e*. gene regulation, signal transduction etc.[2]

Due to the biological importance of protein complexes, it is necessary to develop a specialized database that store, manage and retrieve complex's data effectively.

---

[1] Biochemistry. 5th edition; Berg JM, Tymoczko JL, Stryer L. ;New York: W H Freeman; 2002.

[2] Protein complexes and functional modules in molecular networks; Victor Spirin and Leonid A. Mirny; PNAS 2003 October, 100 (21) 12123-12128.

RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank) is a data resource powered by the Protein Data Bank archive which provides information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

Although there is no doubt in the contribution of RCSB PDB to the biological research, using PDB while research can be involved with some challenges like inconsistent format or protein's unnecessary redundancy.

In our research we are conducting a major revision on a new reliable specialized database based on PDB, which is focused on protein complexes, aimed at simplifying the use of PDB in relation to complex associate research. The main purpose of our research is to add critical computational features which required for publishing the work (complexes database) as paper. Our work includes scraping data from more comprehensive database (*e.g.* PDB) and protein file analysis *e.g.* pdb file format.

Workflow

- **Data collection:**
  The data was collected at previous study. By using advanced search at PDB website[3] the researchers collect 2397 mammalian and bacterial, high resolution heteromers (complexes that their sub-macromolecules are not identical). Moreover, complex's sub-proteins were added.
- **Selection & Cleaning:**
  - **Duplicates elimination:**

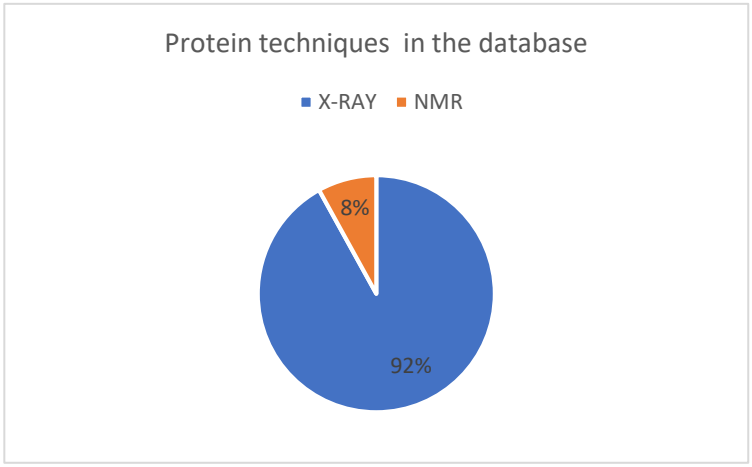
    Sometimes PDB involves high level of redundancy , thus, there are many identical protein complexes, with different names.

    For example:

    The complex 1A0O and the complex 1EAY both contain the same proteins: CHEA and CHEY.
    Our mission was to identify the double complexes and make an organize list of them in purpose to filter the best result of each duplicate complex.
  - **Method – based elimination:**
    Since protein 3D structure determines function, it is indispensable to be able to determine protein reliable structure. Nuclear magnetic resonance spectroscopy (NMR), x-ray crystallography and most recently, cryo-electron-microscopy are three of the most important techniques for elucidating the conformation of proteins. Each method is based on different approach to explore the protein structure, then the structure

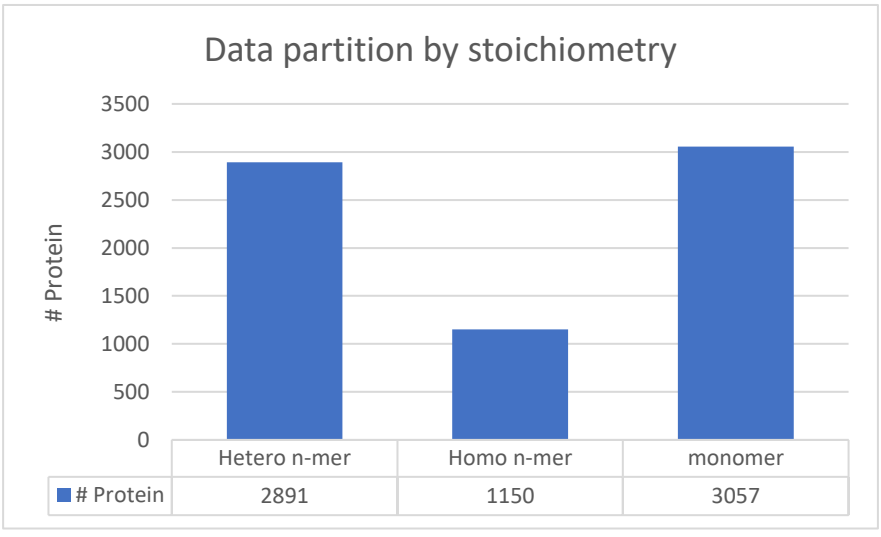
[3] https://www.rcsb.org/pdb/search/advSearch.do?search=new

reliability depends directly on the technique used to determine the structure.
Herein, we classified the data according to the structure detection techniques.



Plot 1 structure detection method

- **Protein stoichiometry:**
  The stoichiometry of a protein complex represents the composition of its subunits. Stoichiometry is useful for predictions about the outcomes of chemical reactions and phenomena explanation.
  We split the data according to the stoichiometry including the subunits, the distribution of the major stoichiometry group is shown in the plot below:



plot 1 Stoichiometry

- **Interface in the complexes:**

  The interface among the subunits in the complex is permanently relates to the complex functionality, thus interface can be a convenient target to drug design for complex inhibition/activation. Crystallographic structures of proteins are usually divided into several fragments which is called "chains". Frequently, the crystallographers provide few identical chains in the same Crystallographic structure.

  In our database, we choose to determine complex interface where the distance between the subunits is under 4.5Å.

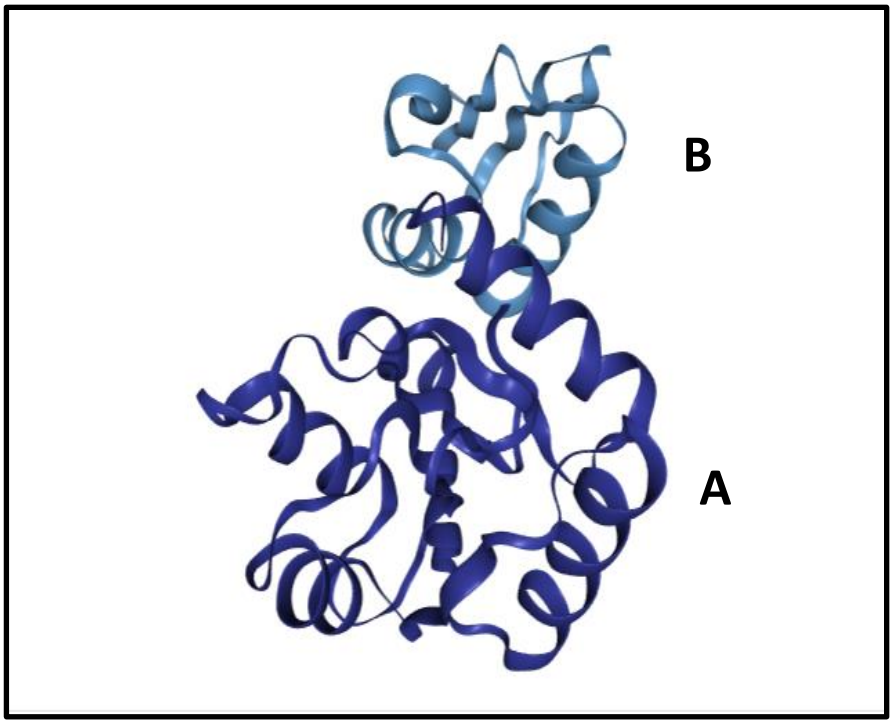


Figure2 The complex 1A0O contains the proteins CHEA and CHEY. CHEA has the chains: B, D, E, F And CHEY has the chains: A, C, E, G. In the complex 1A0O there is contact between chains A and B. In the complex 1A0O there is contact between chains A and B

Sometimes the proteins in the complex contains duplicate chains

So first we had to find the chains of each protein in the complex and remove the duplicate chains. We wrote an algorithm that do it.

For example:

If in protein 1A0O:

CHEA has the chains: B,D,E,F And CHEY has the chains: A,C,E,G

So {'1A0O': [['A', 'C', 'E', 'G']], [['B', 'D', 'F', 'H']]}

After removing the duplicate chains, we got: {'1A0O': [['A']], [['B']]}

That means that in the protein CHEA all the chains are the same, and so are the chains in CHEY.

And after that we found that there is really a contact between the chains: A and B.

As for now, we are working on additional features such as;

- RMSD (*root mean square deviation)* calculation:
  As we mentioned above, the subunits in the complex can also be found as a single protein. Sometimes the protein's experimental structure differs amongst the various conformations.
  Conformation variance in the same protein can indicate about the protein functionality as a subunit versus its function as individual[45].
  At our research we are calculating the level of structural difference between the different protein compositions (as a subunit Vs. individual protein).
- Contact area calculation:
  Thus far, we calculated the distance between the subunits and defined the chains as contact chains if the distance between these chains is less than 4.5Å. The most important - complex's surface area is crucial for developing a PPI (protein-protein interaction). For instance, some studies had found that the proteins binding affinity increase when interfacial surface areas between 500 Å2 and 2000 Å2, the surface energy density decreases as the buried surface area increases. Now, to enrich our database, we are developing a code which calculate the interface area between the subunit.

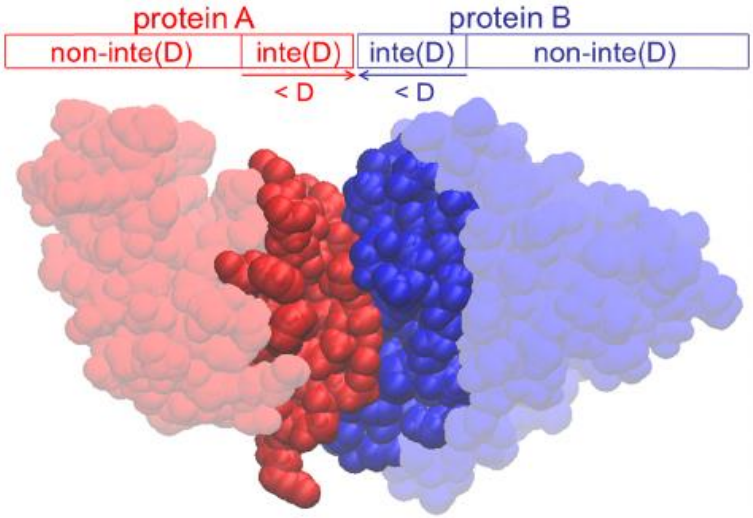


Figure 3 Definition of the interface region, doi:10.1371/journal.pcbi.1003072.g003

[4] ouguet-Bonnet S, Buck M. Compensatory and long-range changes in picosecond-nanosecond main-chain dynamics upon complex formation: 15N relaxation analysis of the free and bound states of the ubiquitinlike domain of human plexin-B1 and the small GTPase Rac1. J Mol Biol. 2008;377:1474–1487.
[5] Simulation Study on Complex Conformations of Aβ42 Peptides on a GM1 Ganglioside-Containing Lipid Membrane; Vahed M, Neya S, Matsuzaki K, Hoshino T, Graduate School of Pharmaceutical Sciences, Chiba University, Kyoto University; Chem Pharm Bull (Tokyo). 2018;66(2):170-177

Therefore, we will update the SQL database, design it and we will send the paper for publishing.

Discussion

As we mentioned above, we examine the data in various aspects that we should combine and decide which records are reliable enough to be in the database. Currently, we did not determine absolute requirements for the database content.

Moreover, in the step of finding interface in the complexes, we had faced some challenges;

Suppose we have a complex of protein 1 and protein 2 where protein 1 contains the chains A and C and protein 2 contains the chains B and D.

If A is similar to C and B is similar to D but A is in contact with D and C is in contact with B, after removing the duplicates we have chains A and B. A and B doesn't make contact, so we can conclude in mistake that there is no contact between chains in the complex.